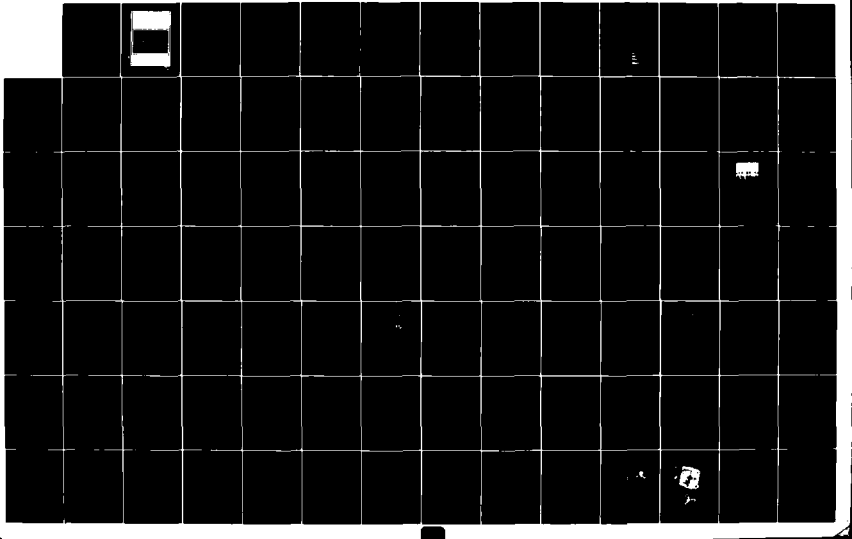
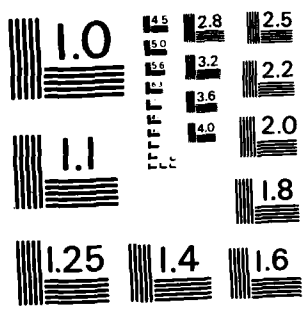
AD-A132 910 ADVANCES IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS(U) ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY.. 1/2

UNCLASSIFIED J L HOLLINGTON JUN 83 AGARD-AG-272 F/G 1/3 NL





AGARD-AG-272

AGARD-AG-272

AG 272

ADVANCES IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CO

AGARDograph No. 272

Advances in Sensors and their Integration into Aircraft Guidance and Control Systems

NORTH ATLANTIC TREATY ORGANIZATION

ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT

(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARDograph No.272

ADVANCES IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS

Edited by

Mr John L.Hollington
T chnical Director
Smiths Industries
Aerospace and Defence Systems Company
Cheltenham Division
Bishops Cleeve
Cheltenham, Glos GL52 4SF
United Kingdom

This AGARDograph was prepared at the request of the Guidance and Control Panel of AGARD

THE MISSION OF AGARD

The mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Exchanging of scientific and technical information;
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Providing scientific and technical advice and assistance to the North Atlantic Military Committee in the field of aerospace research and development;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

Published June 1983

ISBN 92-835-1451-3

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

PREFACE

In the early years of manned flight it was generally believed that the human senses could provide all the inputs to the pilot which he needed to maintain an aircraft in stable flight, but once the performance and endurance of aircraft developed to an extent where extended periods of flight without reference to the outside world became necessary, the severe limits imposed by total reliance upon human sensory perception became apparent. Experimental work between the two world wars led to an understanding of the minimum display of basic flight parameters which was necessary for the proper control of an aircraft in manoeuvring as well as straight-and-level flight, in darkness and bad visibility. Whilst in the early days of "blind flying" the performance of aircraft was such that satisfactory guidance could be provided by relatively simple and usually self-contained instruments, progressive enhancement of performance necessitated improved versions of these instruments, followed in many cases by the development of remote measuring devices (or "sensors"), data transmission systems and slave indicators. These developments provided increased accuracy in the pilot's displays, over a wider dynamic range and in more severe environmental conditions.

In recent years the operational capabilities of aircraft have been extended by the development of new types of sensor. Thus the introduction of inertial references has enabled autonomous navigation systems to be developed, whilst in combat aircraft additional sensors have been introduced which make possible the flight of the aircraft at low level and high speed, the execution of strike and tactical operations at night and in poor visibility, and the recognition, identification, tracking and engagement of targets beyond the visual range of the pilot.

Sensor technology has therefore become as important as the technologies of aerodynamics, materials, structures and propulsion in determining the limits of weapon systems performance and capabilities in military aircraft and is in consequence a major thrust area which attracts a high level of support and activity.

Continuing research and development is leading to enhancement of the capabilities of existing sensors and to the development of fundamentally new types. Furthermore the overall complexity of the total sensor fit in current aircraft and the attendant cost, weight and reliability implications are causing considerable emphasis to be placed on technologies which have the potential to provide benefits in one or more of these areas. Techniques such as analytical redundancy which exploit the multiplicity of data inherent in a total sensor fit with the aim of providing specified levels of redundancy with the minimum total number of sensors are being explored.

The integration of complex modern sensors into the total avionics system is presenting new challenges to the systems designer and manager.

Virtually all the avionics systems and equipment carried by modern military aircraft require sensor data and developments in sensor technology and capabilities are therefore of interest to the systems community as a whole. The purpose of this AGARDograph is to provide the general reader with information on advances in this field and on current approaches to the problems of sensor integration.

Individual papers describe advances in electro-optic sensors and airborne-radar, new types of inertial sensor, low cost fluidic sensors, the measurement of airspeed and windshear with an airborne laser, an application of analytical redundancy and the integration of a new sensor on to an existing aircraft. In addition a paper is included on automatic speech recognition since this rapidly developing technology provides a basis for what may be regarded as a sensor of the human voice which promises a new dimension of possibilities in the man-machine interface.

It is my pleasure to thank many people who have contributed to this publication. I have enjoyed invaluable support from colleagues on the Guidance and Control Panel, and the panel secretariat. I am particularly indebted to Monsieur Bernard Heliot, Guidance and Control Panel Executive for his outstanding administrative support. Finally, but not least, my warm thanks to the individual authors for their papers.

J.L.HOLLINGTON
Editor

CONTENTS

THE STATUS OF LLTV FOR AVIONIC APPLICATIONS

by

W.J.R. Clark
Chief Scientist
Marconi Avionics Limited
Basildon
Essex. SS14 3EL
United Kingdom

SUMMARY

The paper presents an historical review of the development of LLTV cameras and describes the operating principles and salient characteristics of the two types of camera tube suitable for avionic applications. It describes the natural environment in which such systems must operate and the fundamental and practical limitations to their performance. It compares the alternatives and discusses applications experience based upon practical flight trials.

INTRODUCTION

Under the stimulus of technological advances in night vision equipment, ground forces have achieved the capability for effective operation, without supplementary illumination, throughout the hours of darkness. Direct view night sights, low light television and thermal band imaging equipments are all potentially available to the ground forces giving them full night fighting capability and a mobility approaching that of operation in daylight.

In response to this threat, fixed and rotary wing air vehicles are also being provided with an effective means for night-time operation.

All three of the available technologies are potential candidates for fulfilling specialist functions in airborne applications and no generalized assumptions can be made on the matter of choice between them. It is important that the attributes of a particular equipment solution should be recognised and matched to a specific operational requirement and that the practical constraints, of a particular installation, should be considered before a technology choice is made.

This paper deals only with low light television systems operating in, and near, the visual spectrum. It considers image intensifiers only insofar as they may be optically coupled to television camera tubes to achieve quantum limited performance; but not in their more conventional directly viewed form.

The status of thermal band systems has been separately considered in this publication.

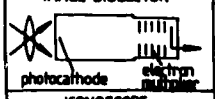
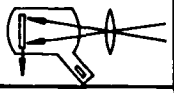
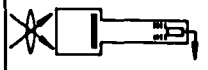
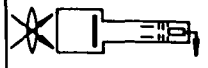
HISTORICAL BACKGROUND

From the origins of electronically scanned television systems in the mid 1930s, until the present time, there has been a steady and continuous improvement in both the sensitivity and image quality of camera tubes. The earlier designs, almost entirely devoted to broadcasting studio application, required high levels of ambient illumination and initially it was the requirement to televise outdoor events that stimulated the development of more sensitive tube types.

These early camera tubes were of the photo-emissive type in which light photons, falling on a photo-sensitive target in vacuo, gave rise to the generation of photo electrons. A variety of electron optical scanning mechanisms were employed to extract the video signals. The essential features of these photo-emissive camera tubes are shown in Fig. 1.

By 1950, the Image Orthicon had emerged as the most commonly used type of camera tube. Subsequently, high sensitivity variants of this design were capable of producing very satisfactory image quality under light levels corresponding to heavily overcast daylight conditions. Whilst producing excellent image quality, the Image Orthicon was physically large and relatively complex, which generally restricted its use to the public service broadcasting field where attended operation, in the hands of skilled technicians was acceptable.

A new type of photo-conductive camera tube, the Vidicon, which first became available about 1950, was very much smaller, more rugged and less complex than the photo-emissive types. In its early forms it was markedly less sensitive than the Image Orthicon, as well as being of lower resolution. Nevertheless, with continued development its performance improved significantly and its availability stimulated the development of compact camera designs suitable for unattended operation for use in industrial and military applications.

NAME OF TUBE	PHOTOCATHODE	STORAGE TARGET	ELECTRON BEAM	VIDEO SIGNAL OUTPUT
IMAGE DISSECTOR 	CONTINUOUS TRANSPARENT PHOTOEMISSIVE CATHODE	NONE	NO ELECTRON GUN	ELECTRONS FROM EACH PICTURE ELEMENT COLLECTED AND AMPLIFIED BY ELECTRON MULTIPLIER
ICONOSCOPE 	PHOTOEMISSIVE TARGET WITH INSULATED ELEMENTS DEPOSITED ON MICA SHEET BACKED BY CONDUCTING SIGNAL PLATE		HIGH VELOCITY ELECTROSTATIC FOCUS	VIDEO SIGNAL INDUCED IN TARGET SIGNAL PLATE
ORTHICON 	PHOTOCATHODE & TARGET COMBINED IN SIMILAR MANNER TO ICONOSCOPE EXCEPT TARGET SIGNAL PLATE IS TRANSPARENT		LOW VELOCITY MAGNETIC FOCUS ELECTROSTATIC DEFLECTION	VIDEO SIGNAL INDUCED IN TARGET SIGNAL PLATE
IMAGE ORTHICON 	CONTINUOUS TRANSPARENT PHOTOEMISSIVE CATHODE	THIN SEMICONDUCTING GLASS TARGET WITH FINE MESH SCREEN TO COLLECT SECONDARIES	LOW VELOCITY MAGNETIC FOCUS & DEFLECTION	MODULATED RETURN BEAM PASSED INTO ELECTRON MULTIPLIER
IMAGE ISOCON 	CONTINUOUS TRANSPARENT PHOTOEMISSIVE CATHODE	SAME AS IMAGE ORTHICON	LOW VELOCITY MAGNETIC FOCUS AND DEFLECTION WITH SCATTERED ELECTRON SEPARATOR	SCATTERED FRACTION OF MODULATED RETURN BEAM PASSED INTO ELECTRON MULTIPLIER

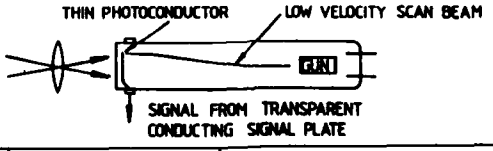
TYPES OF PHOTOEMISSIVE CAMERA TUBE

FIGURE 1

The early Vidicon camera tube employed an Antimony Trisulphide target scanned by a low velocity electron beam. Subsequent development of improved targets such as (Chalnicon, Newvicon or Saticon) Lead Oxide (Plumbicon or Leddicon) and the silicon diode array target (Sidicon), together with refinements in electron-optical design provided further improvement upon the performance of the basic Vidicon; until today when they have emerged as the primary sensors for virtually all daylight performance cameras. Fig 2 reviews the essential characteristics of Photo-conductive camera tubes. Faced with a growing military demand for equipment capable of operating by both day & night, under naturally occurring illumination, the last two decades have seen the development of variants of both the Image Orthicon and the basic Vidicon, to provide substantial improvements in sensitivity.

In the case of the Image Orthicon, the design approach was to further develop the photocathode and target sensitivity and, by means of a novel method of reducing the return beam current, to substantially reduce thermal noise components originating in the beam. This specialised low light television tube is known as the Image Isocon.

Two highly sensitive camera tubes, employing a derivative of the Vidicon electron gun and scanning mechanism similarly evolved. Firstly, the Secondary Electron Conduction camera tube (SEC) and subsequently, with the development of silicon targets, the Silicon Intensifier Target tube (SIT or Ebsicon). In both these cases the high sensitivity has been obtained by the introduction of an image section within which photo electrons originating at a photocathode are accelerated under the influence of a high potential to arrive at the target layer with high energy. As a result, electron multiplication in the target layer provides a signal gain mechanism. In both cases the signal scanning mechanism remains substantially the same as their progenitors.



NAME OF TUBE	PHOTOCONDUCTOR MATERIAL	TARGET CONSTRUCTION	GENERAL CHARACTERISTICS
VIDICON	$Sb_2 S_3$	THIN POROUS PHOTO-CONDUCTOR ON TRANS-PARENT CONDUCTING SIGNAL PLATE	GOOD DAYLIGHT SENSITIVITY. PERCEPTIBLE LAG AT LOWER LIGHTING LEVELS γ = 0.7 HIGH RESOLUTION EASY TO MANUFACTURE. GAIN CONTROL BY VARIATION OF TARGET VOLTAGE.
LEAD OXIDE VIDICON	Pb O	APPROXIMATELY SAME AS VIDICON	HIGHER SENSITIVITY THAN VIDICON BUT LOWER LAG. GOOD SIGNAL UNIFORMITY. γ = 1.0 GOOD RESOLUTION
SILICON VIDICON	Si	ARRAY OF p - TYPE PHOTODIODES ON THIN n - TYPE SILICON WAFER	HIGHER SENSITIVITY. LOW LAG. RESISTANT TO BURNED-IN IMAGES RESOLUTION LIMITED BY PHOTODIODE STRUCTURE. SENSITIVE TO NEAR IR. MORE COSTLY THAN VIDICON
CADMIUM SELENIDE VIDICON	Cd Se (+ ZnS or $AS_2 S_3$)	CONTINUOUS DOUBLE LAYER	HIGHER SENSITIVITY THAN VIDICON. GOOD RESOLUTION LOW LAG GOOD SIGNAL UNIFORMITY
SIT VIDICON	S20 / S25	ARRAY OF PHOTODIODES ON SILICON WAFER	VERY MUCH HIGHER SENSITIVITY THAN VIDICON DUE TO TARGET GAIN MECHANISM. MODERATE RESOLUTION

TYPES OF PHOTOCONDUCTIVE TUBES

FIGURE 2

A parallel development during the 1960s, having considerable impact on the subsequent design of low light T.V. systems, was the Direct View Image Intensifier. In its first generation form these devices consist of a vacuum tube having a photocathode at one end and a display phosphor screen, deposited on a fibre optic plate, at the other. Photo-electrons, emitted from an optical image formed on the photocathode, are accelerated and focused at high potential to form an electron image on the display phosphor. Optical image gains of X 40 are typical of single stage Direct View Intensifiers of this type. By coupling a Direct View Intensifier to the new forms of high sensitivity camera tubes, it was now possible to generate television images under light levels at which the imaging performance was determined by the particulate nature of light itself. The quality of the image is therefore determined by the number of photons available to form the image in the television field scan period. In this photon limited situation, further improvements in the basic camera gain mechanisms becomes unrewarding. Only by improvements due to the attainment of higher quantum efficiency photocathodes and by image processing techniques, is further improvement in system sensitivity likely to be realised.

More recently, second generation devices, based upon electrostatically self-focusing micro-channel plate image intensifiers, can provide an optical gain in excess of 65000. If these highly sensitive intensifiers are coupled to a simple photo-conductive camera tube, photon limited perf-ormance can also be achieved; this combination is becoming quite widely used for the less exacting applications where low cost and size, rather than performance, are the primary concerns. Third generation intensifiers using Ga As photocathodes with improved quantum efficiency photocathodes will be in use within the next few years.

OPERATING PRINCIPLES

In the military applications of LLTV, the history of competition between photo-emissive and photo-conductive camera tubes has been repeated. The essential choice lies between achieving the best available performance on the one hand and lower equipment weight, size and sometimes cost on the other. Of the available alternatives, only the Image Isocon and the Ebsicon have been used to any extent in avionic LLTV applications and this review will concentrate on discussing the characteristics and applications of these tubes.

a) The Image Isocon.

Figure 3a shows a diagramatic sectional view of an Image Isocon with the camera located within its associated electro-magnetic scanning and focusing coils. Light from the scene to be viewed is imaged by a lens onto a photocathode on the inner surface of the faceplate.

As shown in the diagram, if the camera tube is to be coupled to an Image Intensifier, a fibre optic faceplate is used, rather than a plane glass window which would otherwise be employed. The optical image formed on the photocathode gives rise to the release of photo electrons, which are accelerated and focused by electro-static and magnetic fields. Passing through the target mesh, they strike a conducting glass target, releasing secondary electrons. The target mesh is held a few volts positive relative to the thermionic cathode so that secondary electrons are captured, to leave the target with a positive charge pattern 'image' impressed upon it. At the high density of secondary electrons corresponding to peak white areas in the scene, the target charge pattern approaches the target mesh potential, thus tending to limit further capture of electrons by the target. However, some of these secondary electrons in whiter parts of the scene do, in fact, fall back on the target, selectively enhancing the positive charge image. This gives pictures derived from an Isocon a characteristically 'crispened' appearance. It is this dynamic interaction between the charge image and the mesh potential which largely determines the transfer characteristic of this type of tube and its excellent overload performance.

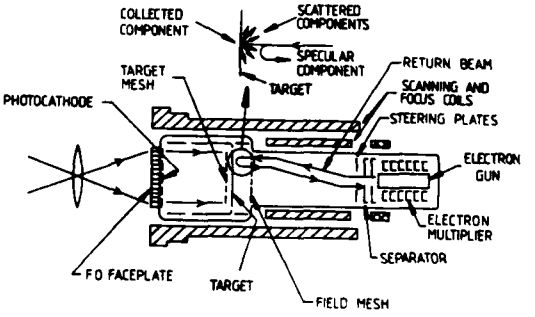


FIGURE 3a

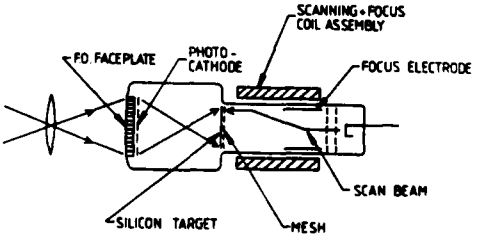


FIGURE 3b

Turning now to the scanning beam section. The target, being made of a special conducting material, allows the positive charge to transfer to the rear surface where it can be scanned by a magnetically deflected electron beam.

The scanning beam, due to the decelerating influence of a field mesh, approaches the target uniformly at near zero velocity, landing on the target only in those regions where the target is more positively charged. In areas of the target corresponding to the darkest parts of the optical image there is little or no charge and the beam fails to land. It therefore returns back down the tube virtually retracing the path of the original electron beam. In this way, the

return beam has become modulated with a beam current component which is an analogue of the charge pattern along each scanning line, (and hence, of the original scene brightness fluctuations).

In the Image Orthicon, this return beam current is allowed to fall on the first dynode of an electron multiplier stack, at the final anode of which the signal current flows.

In the Isocon mode of operation the signal recovery mechanism is subtly different. When the scanning beam reaches the target, the return beam splits into two components. One, the specularly reflected beam we have already described and the other, a beam formed of scattered electrons. In effect the 'scattered' beam represents the small proportion of the original beam which is neither collected by the target, nor specularly returned. These scattered electrons exhibit the special characteristic that they are directly proportional to that part of the beam which actually lands on the target. Therefore they carry only picture information. It will be noted that in dark parts of the charge image, where no electrons land, there is no scattered beam and hence no shot-noise.

By contrast the specular beam, which contains most of the original scanning beam, contains a high shot-noise component, particularly in areas corresponding to dark parts of the scene.

In the Image Isocon a steering plate is used to impart a transverse component of velocity to the original beam so that it follows a helical trajectory as it passes down the tube. Because of this transverse component, the scattered return beam electrons follow a slightly different trajectory from the specular components and this enables a separator plate to be used to intercept the specular beam, but allow the scattered beam to fall on the first dynode of the electron multiplier. Compared with the Image Orthicon, the much lower shot noise component in the Image Isocon's return beam permits additional stages to be employed in the electron multiplier stack. In effect, the electron multiplier provides a relatively noise free gain mechanism, increasing the signal current to a level at which subsequent noise contributions from the camera head amplifier are insignificant.

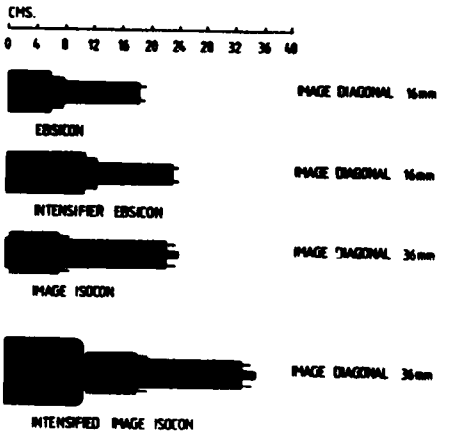
b) Silicon Intensifier Target Camera Tube (Ebsicon).

A schematic diagram of an Ebsicon is shown Fig. 3b. The input window, usually in the form of a fibre-optic faceplate to facilitate coupling to a single stage of optical image intensification (not shown), has a photocathode deposited on its inner surface. When an optical image is formed on the photocathode, photo electrons are emitted which are accelerated and electrostatically focused under the influence of high potentials to form an electron image on a thin silicon wafer target upon which a very tightly spaced matrix of p-n junctions have been formed. The spacing of the diodes is of the order of 14µm. A gain mechanism is caused because the photo electrons arrive with an energy of up to 10KeV which causes multiple dissociation of the electron-hole pairs. The holes are collected at the p-side of the diodes where the charge is neutralized by the scanning process of the electron beam. In effect, as the beam scans across the diode array, a signal current is read out from the baseplate of the target electrode. Typically the signal in white areas of the scene is 300nA.

The high sensitivity of the SIT Vidicon is a consequence of the energy imparted to the photo electrons and target gains of up to 2000 are claimed. The gain may be varied by a factor of up to 100:1 by reducing the EHT supply, but operation at maximum gain over the full control range does incur performance penalties. In a practical design, gains of 1000 and control ranges of 30:1 are more typical.

With a single stage of image intensification preceding the photocathode, photon limited performance can be achieved. Typically the intensifier will also be operated at a potential of of about 10Kv and it is usual to control both the intensifier and the image section to vary the overall gain of the combination. In this way gains in excess of 30,000 are achievable with a maximum gain range of 8000:1.

The Ebsicon camera tube is thus relatively simple in operation, rugged and has a gain mechanism which is readily controllable by electronic means



LLTV CAMERA TUBE SIZES Figure 4

SENSOR CHARACTERISTICS

A comparison of the saliert physical characteristics of the two classes of LLTV camera tubes most commonly used in avionic applications is shown in the scale drawings of Figure 4.

Both the Ebsicon and the Image Isocon can be coupled to a single stage intensifier to enhance their sensitivity and achieve photon limited performance. In particular it should be noted that the Image Isocon photocathode diagonal (36mm) is greater than the Ebsicon (16mm). As a result the photocathode area is greater by a factor of 5:1. Thus for objective lenses having e same f/No the number of photons contributing to the image is greater in proportion to the image area. This confers a significant sensitivity advantage to the Isocon but at the expense of a larger optical aperture. Perhaps surprisingly, the maximum diameter of the two types of camera tube is substantially the same. One reason for this is that the high voltage associated with the Ebsicon image section requires that the front section of the tube is encapsulated to withstand the avionic environment.

TYPICAL SPECTRAL RESPONSE.

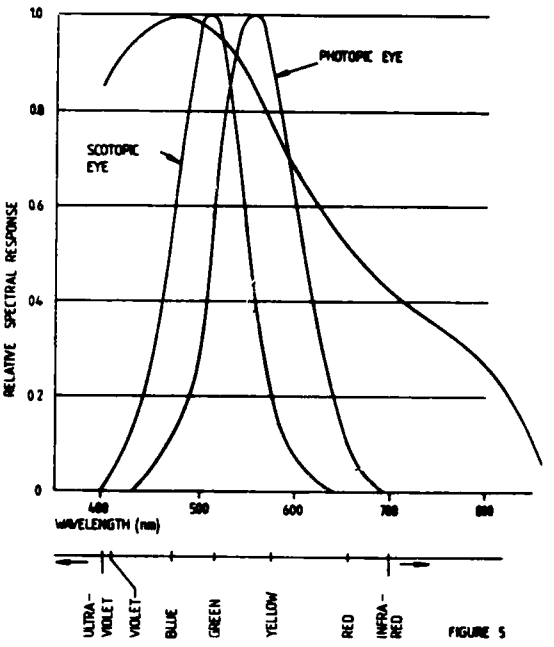


FIGURE 5

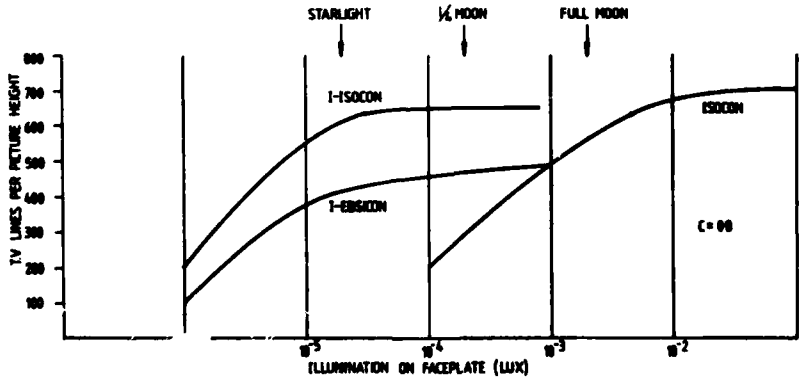
The photocathode response of the two types is substantially similar and a typical characteristic is shown in Figure 5 where it is also compared with the human eye operating in its photopic (daylight) and its scotopic (night time) regimes. The considerably enhanced response of the photocathode to radiation at the red end of the spectrum is evident from this data. These LLTV photocathodes have been specially developed to have this high sensitivity in order to obtain maximum benefit from the high energy content of natural illumination at the longer wavelengths.

The imaging performance of LLTV camera tubes are most commonly defined in terms of their resolution capability as a function of the illumination incident upon the camera tube photocathode. The term 'resolution' describes the ability of the camera to present a visual image of a finely structured pattern and, by common usage, the chart most widely adopted for assessment purposes is a regular vertical pattern of equal width black and white bars. Traditionally tne bar pattern is defined in terms of TV lines/picture height. That is, the sum of the number of black and white bars that could be fitted into a dimension equal to the height of the image.

Figure 6 shows a typical set of static resolution curves for the Isocon and both the Isocon and Ebsicon camera tubes fitted with 1:1 intensifiers. Resolution is expressed as a function of the illumination falling on the camera tube faceplate measured in Lux (lumens/m^2). These curves have been measured with test charts having a black to white contrast of 0.8. Curves of this type are also known as the 'limiting resolution' characteristic since each measurement is made by noting the highest spatial frequency (lines/picture height) at which the pattern is just visible in the displayed image.

The relationship between scene illumination and faceplate illumination depends, of course, on the lens aperture and its transmission. Figure 6 also shows the corresponding natural illumination levels based upon use of an f1.4 lens.

This particular set of curves were all measured by the same methods and in representative camera equipments of military standard. They therefore represent a reasonable basis for comparison. They show the higher resolution obtainable from the Isocon and the high sensitivity achieved by both types of tube when fitted with an intensifier.



STATIC RESOLUTION CHARACTERISTIC

FIGURE 6

In both cases 1:1 magnification intensifiers were fitted; the Ebsicon with a 16mm diagonal image format and the Isocon with a 36 mm diagonal. It is of course possible to increase the sensitivity still further by utilizing a larger input photocathode and employing a de-magnifying intensifier to match to the camera tube photocathode dimensions. For example: flight trials have been carried out with a 2:1 intensifier fitted to an Isocon. This combination probably represents the best that can currently be achieved with LLTV and represents about a 4:1 sensitivity improvement over the I-Isocon curve in Figure 6. However, such a system is probably too bulky for all but the most specialist applications.

In extrapolating from this type of measured camera data to predictions of performance in a practical application, it is important to recognise the curves tend to present an over optimistic result. In the first place, real scenes are of lower contrast than the test charts typically used in the measurements. The effect of reduced scene contrast is to lower the effective sensitivity; typically by a factor of about 5:1 for very low contrast subjects. Secondly, the curves are for static scenes and, because all currently available camera tubes integrate the incident radiation between successive frame scans, they are subject to image blur. In effect this can be considered as a loss of resolution due to image motion. In avionic applications, with a well stabilized system, the effects can be quite negligible on distant objects or when a point on the scene is being tracked effectively reducing relative motion effects. However, with foreground objects, or during periods of aircraft manoeuvre, the effects are not negligible and then an effective loss of system resolution must be expected.

As a general conclusion it can be stated that the best available systems, whether I-Isocon or I-Ebsicon, can produce satisfactory results down to clear starlight levels for all but the lowest contrast scenes, and useful results below this on higher contrast subjects, providing image crossing rates are low.

It is beyond the scope of this review to consider the many other more specialised parameters of LLTV cameras which are used to characterize performance but brief mention must be made of noise since, in the limit, it is this parameter which dominates achievable performance. At the higher light levels, noise generated within the tubes, and in the case of the Ebsicon also in the head amplifier, represents the main sources. The noise mechanisms in the two types of tube is in fact quite different. As we have seen in an earlier section, because the Isocon noise originates primarily as shot noise in the beam, which tends towards zero in dark areas of the scene, a picture is produced which is relatively noise free in the darker areas but which increases towards peak white. This is a very desirable characteristic since the eye is more tolerant of white noise than black. In the Ebsicon the dominant noise components originate in the target gain mechanism with added contributions from the head amplifier. It is thus distributed throughout the grey scale and dependent on target gain and hence on light level. At the higher light levels the Isocon undoubtedly exhibits superior performance and subjectively produces relatively noise free pictures.

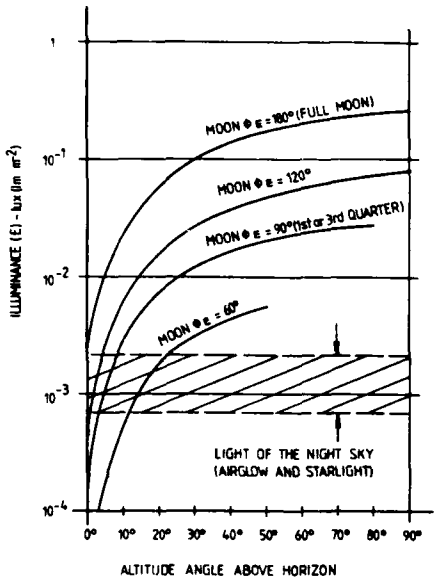
For both tubes as the light level is reduced, signal noise level falls and photon noise increases. In the limit the photon noise becomes dominant and for most practical purposes it is this parameter, rather than the falling resolution, which sets the lower performance limits of both tubes. It is always difficult to define the exact point at which an image becomes too noisy to be usable to accomplish a specific task, because it is so dependent on the scene contrast and structure. In practical terms, there is a progressive degradation below a certain level which, for either tube fitted with an intensifier and for the best achievable lens aperture, is in the clear starlight region. Under these conditions, a side by side comparison invariably shows the Isocon to be superior but it must be remembered that this is mainly due to the larger format with consequent increase in the lens aperture. For two systems constrained to have the same lens diameter, the Isocon shows less advantage.

The rather subjective discussion of 'performance' in this section has aimed to overview the topic and indicate the more important factors underlying design choice. The scientific prediction of performance is discussed in a later section on mathematical modelling.

THE OPERATIONAL ENVIRONMENT

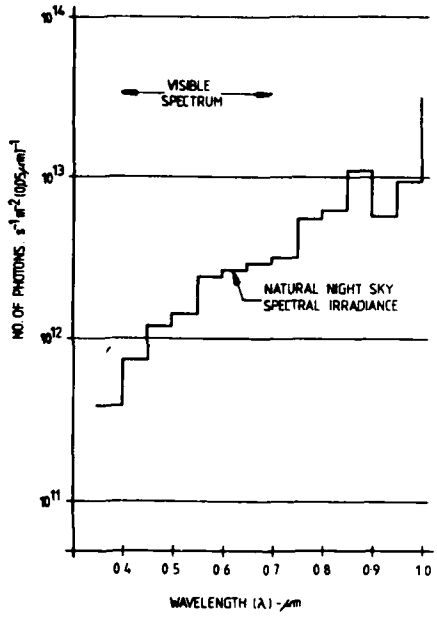
There are two sets of natural phenomena which together determine the performance of an LLTV system. The first of these is the natural irradiance of the scene being viewed and the second the optical properties of the atmosphere that lies between the LLTV camera lens and the scene. To a lesser extent, the optical properties of the atmosphere also influence the natural irradiance on the scene. In general these two phenomena are largely independent insofar as the first is strongly time dependent and predictable (i.e. with the diurnal, lunar and seasonal cycles), whereas the second is mainly influenced by the prevailing air mass, depending on large scale meteorological phenomena in a much less predictable way.

In considering natural irradiance it is convenient to confine our attention to the hours of darkness. The moon has to be considered apart from the other sources since its influence can range from being dominant to being negligible. In each 28 day period, for some 4 - 5 days on either side of new moon, its effect can be discounted for practical purposes both because its apparent surface subtends a negligibly small area and also because it appears above the horizon only near twilight when scattered sunlight is still the significant source irradiance. On the other hand, near full moon, the source is large and it is highest in the sky near midnight. The effect of these changes and the levels of illumination relative to the other sources is shown Figure 7.



ILLUMINANCE LEVELS ON THE SURFACE OF THE EARTH DUE TO THE MOON AND THE SKY.

FIGURE 7



NATURAL NIGHT-SKY SPECTRAL IRRADIANCE ON HORIZONTAL EARTH'S SURFACE

FIGURE 8

In the absence of moonlight, the sky's irradiance originates from, and is distributed between, the following sources:-

Zodiacal Light	15%
Galactic	5%
Luminescence of the night sky (variable)	40%
Scattering from the above	10%
Direct & Scattered starlight	30%
Extra galactic sources	< 1%

From a spectral point of view, the moonless night sky itself has a marked red and infrared predominance as shown in Figure 8. One should note however, that the ordinate of this diagram is in units of photons/sec. When converted to watts, the bias towards the red is rather less pronounced.

Moonlight is spectrally very similar to sunlight, (approximating to a black body at 5900°C) since the moon's surface is nearly spectrally neutral as a reflector.

Figure 9 summarises the naturally occurring levels of illumination. An ideal LLTV system would operate effectively over the range some 5 orders of magnitude below twilight conditions (10 lux) but as shown elsewhere in this review, practical systems fall short of this ideal.

In the foregoing discussion, only the natural sources of irradiance have been considered and quantified. In the real world however, man-made sources of illumination can have significant influence, particularly toward the lowest end of the irradiance scale. The effects of scattered lighting from large conurbations is particularly marked when there is continuous cloud cover, and indeed the darkest conditions may occur under clear starlight conditions. For this reason in peacetime for example, light levels rarely fall much below 10^{-3} lux in South East England.

Sky Condition	Approximate Levels of Illuminance - Lux (lm m^{-2})
DIRECT SUNLIGHT	1-13 x 10^5
FULL DAYLIGHT (Not direct sunlight)	12 x 10^4
OVERCAST DAY	10^3
VERY DARK DAY	10^2
TWILIGHT	10
DEEP TWILIGHT	1
FULL MOON	10^{-1}
QUARTER MOON	10^{-2}
MOONLESS, CLEAR NIGHT SKY	10^{-3}
MOONLESS OVERCAST NIGHT SKY	10^{-4}

NATURAL SCENE ILLUMINANCE.

Sky Condition	Approximate Values of Luminance - nit (cd m^{-2})
CLEAR DAY	10^4
OVERCAST DAY	10^3
HEAVILY OVERCAST DAY	10^2
SUNSET, OVERCAST DAY	10
1/4 HOUR AFTER SUNSET, CLEAR	1
1/2 HOUR AFTER SUNSET, CLEAR	10^{-1}
FAIRLY BRIGHT MOONLIGHT	10^{-2}
MOONLESS, CLEAR NIGHT SKY	10^{-3}
MOONLESS, OVERCAST NIGHT SKY	10^{-4}

FIGURE 9

APPROXIMATE VALUES OF LUMINANCE OF THE SKY NEAR THE HORIZON

The state of the atmosphere has only a relatively small influence on the light level derived from extra terrestial sources but can have a very profound influence on the apparent contrast of objectives within the scene. In the visible and near IR parts of the spectrum the normal atmosphere, including water vapour, has only a small effect on contrast except at very long ranges. Under very clear conditions it is molecular scattering which sets an ultimate limit to contrast and hence to visibility.

The water droplet content of the air is, however, a crucial factor in the appearance of a distant scene. The main influence is due to scattering of light and, except perhaps in industrial regions, absorption can be neglected. The effect of scattering is a reduction in contrast which is approximately exponential with range and it is almost always this factor which sets a limit to the range at which an object can be viewed (assuming of course, line of sight to the object and an adequate sensor resolution). Figure 10 gives an indication of the probabilities of occurrence of various visual ranges. The data relates to the N.W. German plain (the statistics vary of course, with geographical location). Also, most of the available data relates to visibility measurements made during the daylight hours. There is evidence to suggest that night time visibility is less, but because such measurements are more difficult to obtain, statistical data is not yet widely available. Figure 10 should therefore be considered an optimistic assessment if applied to LLTV systems.

As a rather general statement it is true to say that as far as the effects of visibility are concerned, an LLTV system suffers from much the same effects as the eye. However, it has been shown that electronic contrast enhancement can result in markedly superior results when the signal/noise ratio is sufficiently good to permit this form of processing. Of course the LLTV system always gains over the eye at the lower irradiance levels where the eye's acuity and sensitivity are no longer adequate to 'see'the scene.

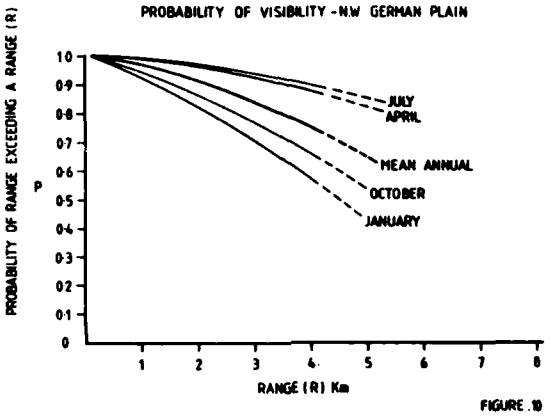


FIGURE 10

FUNDAMENTAL LIMITATIONS

Whatever the degree of technical sophistication that is incorporated into an LLTV camera tube or its associated electronics, a fundamental limit to picture quality is set by the information arriving on the photocathode from the scene itself. The radiometric information is contained in the average numbers of photons arriving from various parts of the scene. The visibility of a particular target against its background will depend on the difference between the appropriate number of photons arriving at the input aperture during the integration time of the sensor (i.e. the television frame rate).

The number of photons arriving in unit time are however, subject to statistical fluctuations and this, in turn, gives rise to fluctuations or noise in the reproduced luminance on the display. This 'photon noise' leads to a fundamental limit to the performance of any electro-optical imaging device which relies on natural irradiance of the scene. When the display surface is viewed by an observer, the human eye-brain system tends to integrate areas of the picture in order to discern differences in their average luminances. The larger the integrated area, the smaller are the perceivable fluctuations in its average luminance. Thus small luminance differences are more likely to be perceived between large areas than between small areas. Therefore the resolution of detail falls as the noise increases.

The rms fluctuation in photon arrivals is equal to the square root of the number of photons involved and hence, so also is the signal/noise ratio. Thus as the light level falls, so does the S/N ratio associated with a given luminance difference.

Expressed in subjective terms, as the light level is decreased, the noise level increases, the contrast between an object and its background falls and the resolution falls. Under these limiting conditions the system is said to be photon limited.

PRACTICAL LIMITATIONS

For a photocathode of given quantum efficiency, the natural irradiance level at which photon limited performance occurs will be influenced by the integration time and the achievable optical aperture (or light collecting power of the lens). Increasing the integration time beyond the normal television frame period is impractical in most avionic applications, since this also results in reduced dynamic performance (image blur due to air vehicle motion). A practical limit to the light collecting power of the lens, in many avionic applications, if often set by the diameter of the optical system that can be accommodated. For wide angle systems (short focal length), the relative apertures are small and in such cases the photon limited resolution threshold can be lowered by utilizing a sensor with a large photocathode diameter (36mm or even larger). However, narrow angle systems have to employ optics which are almost always 'diameter' limited by installation restrictions. In such cases there is not much to choose between a large photocathode and a relatively high 'f'/No lens - or a smaller photocathode and a smaller 'f'/No lens.

From these considerations it is evident that each particular application should be carefully designed to obtain the best possible quantum limited performance achievable within the practical constraints of the volume available.

PERFORMANCE MODELLING

Reliable performance predictions are an essential part of electro-optical system development. The designer is faced with a relatively complex series of trade-offs which usually involve cost, weight and size versus performance. Frequently the system must be integrated with other avionic functions, and this may require, for example, target acquisition range to be matched to the weapon's characteristics. However, electro-optical systems usually involve a human observer and operate in a natural environment which, as we have seen, is quite complex and variable. It is therefore not surprising to find that simple deterministic performance predictions are not possible. Similarly, it is not practical to obtain a comprehensive assessment of performance by field trials owing to the multi-dimensional set of variables that have to be covered and the statistical nature of some of the parameters. A much cheaper and very much more flexible alternative is to employ computer simulation. In recent years considerable R & D effort has been devoted to deriving credible mathematical models.

Computer modelling techniques have been applied to every aspect of the problem, from the natural scene itself, to the eye and brain of the observer who views the final display. Such models treat in considerable detail the spectral aspects of target surfaces, the natural terrain backgrounds, the atmosphere and the complications of irradiance, both natural and artificial. The optical system and various types of camera tubes have been modelled including the Ebsicon, Isocon and Vidicons, as well as Image Intensifiers. Electronic processing, such as contrast correction, filtering and various contrast enhancement techniques are usually included as well as the display itself.

A concept, usually referred to as 'perceived signal-noise', is frequently introduced to account for properties of the human eye and to arrive at probabilities of detection, recognition and identification.

A mathematical model has little real value as a system design tool until there is reasonably high confidence in its accuracy. Considerable effort has therefore, been devoted to its validation. Large parts of the model have been tested by comparing direct measurement with the predicted response to a known stimulus. An interesting consequence of the iterative process of modelling and validation in this area has been the extent to which there is now an improved understanding of the product being modelled. Indeed, a good model can serve as most valuable equipment design tool as well as fulfilling its original function of predicting system performance!

In the case of the observer/display interface the validation is more difficult and confidence can only be built as the result of the numerous experimental trials that have been carried out throughout the electro-optical community. One particularly difficult area which is not yet fully resolved concerns task accomplishment in the presence of 'clutter'. A particular target may be clearly perceived once it is found (or pointed out) but the psychologically disturbing presence of 'clutter' introduces an uncertainty in the search time that may be necessary.

Despite these reservations, experience in exercising the whole model has shown very good agreement between predictions and trials results. As a result there is now a general acceptance of the validity of models and relevance on performance predictions produced in this way. Fig 11 shows a set of observations carried out on a particular trial of an LLTV system and a probability of detection curve produced by a mathematical model representing the measured trial conditions.

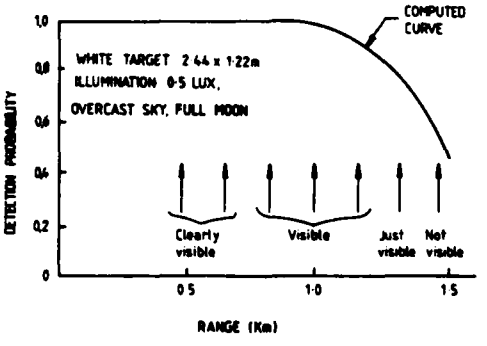


FIGURE 11 COMPARISON OF MODEL WITH EXPERIMENT

APPLICATIONS EXPERIENCE

Practical experience gained from extensive trials of LLTV equipment based on both Ebsicon & Isocon camera tubes has shown that the results achieved correlate well with modelling predictions. There is now a wealth of experience available from flight trials to show that aircraft can be flown safely, at low altitudes under illumination conditions extending down to clear starlight or overcast ¼ moonlight. For fixed wing aircraft piloting applications, a Head Up Display (HUD) presentation is preferred with the T.V. image overlaid on the real word with a 1:1 magnification. In this application the widest possible field of view is necessary to enable the pilot to see into turns but in practice, a limitation is imposed by the angles achieved in the HUD optics. In this respect the new holographic HUD designs offer a significant advance over earlier displays and a 30° horizontal angle is now achievable. With such wide angles the camera lens focal length is short and the physical diameter, at the best attainable numerical aperture, is acceptably small. In this application the Image Isocon is the clear choice because of its superior resolution and generally low noise image quality.

For targeting systems or air-to-air identification aids, a high magnification optical system is invariably required to achieve long range performance. In these cases, because long focal length lenses are required, physical installation constraints severely limit the lens aperture that can be used in avionic systems. This means that the large photocathode advantage of the Isocon cannot be exploited and the smaller photocathode Ebsicon tends to be the favoured solution. Most of the practical experience with long range viewing systems is therefore based upon Ebsicon camera equipment.

As is so often the case with practical trials experiences, many lessons have been learned which were not altogether expected, although perhaps obvious with hindsight. For example, experience has shown that the apparent quality of the image from a flying aid LLTV camera improves markedly as altitude is reduced. In part this is due to the fact that terrain features are more clearly resolved at the lower altitude but more particularly that up-standing features such as trees, buildings and hedges, when seen in side elevation, present better contrast, perspective and streaming cues than when seen from higher altitudes. This is an unexpected but fortunate effect which greatly improves pilot confidence in low flying by night.

Practical experience has shown the importance of reducing specular reflection in the region surrounding the lens and the fact, for example, that inclined plane windows are superior to spherical window surfaces in minimising the effect of off axis sources of illumination (such as the moon).

Another not insignificant problem encountered in practice, is the influence of bright lights within the viewed image. It should be remembered that on a dark night, man-made light sources may be 4 or 5 orders brighter than the surrounding scene. From measurements and trials it has been shown that the Isocon camera, properly designed to minimise overload effects, is superior to the Ebsicon by a factor of some 40:1 (defined as the ratio of intensities for the same overload enlargement effect) and this becomes an important factor influencing the choice of this type of tube as a flying aid. With a well designed system the influence of clusters of lights is reduced to a nuisance value and is not now considered as serious a problem as was once the case.

CONCLUDING SUMMARY

As a result of considerable development during the last decade and extensive flight trials, notably those carried out by the Royal Aircraft Establishment, LLTV has emerged as a practical option for avionic use. Equipments have been designed which, with reservations concerned with the upper temperature limit, are capable of meeting avionic environmental specifications and are adequately stable for service use. To this extent the technology can be considered mature and available for practical application.

Its main disadvantages lie in its dependence on atmospheric visibility which affects displayed contrast in much the same way as it does to the human eye, and a photon limited performance at the very lowest light levels.

In many respects LLTV systems emerge as being complementary to thermal imaging with a superior performance in rain and under conditions of thermal wash out but inferior performance in mist or through smoke. Because LLTV operates in much the same spectral band as the eye, its images present no interpretation difficulties but on the other hand, it lacks the ability to differentiate between natural and man-made (heat producing) objects as a thermal imaging system can.

The ideal Electro-optical system would include both technologies to capitalize on their complementarity, but this is unlikely to be an economically viable proposition in any but the most specialized applications.

In the final analysis, the choice will be made on the basis of the detailed requirements of the application, on performance, cost and practical considerations such as logistic support requirements. It is perhaps fortunate that we now have two techniques ready and able to meet the need for night operations where, until recently, we had none.

ACKNOWLEDGEMENTS

The Author wishes to thank the originators of the many sources of information used throughout this review.

THERMAL IMAGING FOR AVIONIC APPLICATIONS

by

G.M. Cuthbertson

Chief Systems Engineer

Electro Optical Surveillance Division

Marconi Avionics Limited

Basildon.

SUMMARY

The quality of visual images formed from thermal radiation is now of a standard which permits thermal imagers to be widely used in military avionic systems. The paper briefly outlines the physical principles underlying the technology and considers the natural environment in which it operates. It briefly traces the evolution of thermal imager designs and describes the way that the latest advances in detectors and circuit technology have been implemented in the Class II U.K. Thermal Imager Common Module (TICM II) programme.

INTRODUCTION

In the last decade the quality of visual images formed from the thermal radiation emitted by objects at or near ambient temperature has progressed to the point where they are virtually indistinguishable from the best monochrome television pictures. Thermal imagers have achieved the standards of thermal and spatial resolution necessary to fulfil a range of military functions and their engineering development has progressed to a point where reliable field use is becoming widespread. Since no illumination is required, the performance of a thermal imager is virtually independent of the time of night or day and their short range performance is independent of most weather conditions. Even at longer ranges, where atmospheric absorption can become significant, utilization factors are invariably better than other forms of visual or electro-optical sensor.

The military implications of thermal band imaging systems are clear. They permit wide angle viewing systems which are rarely weather limited, to be designed to aid the piloting or night driving task. When fitted with a high magnification lens they enable medium to long range detection, recognition and tracking of military targets. They are passive and substantially unaffected by battlefield smoke, haze and dust.

This paper will briefly outline the physical principles underlying the technology and then consider the natural environment. After tracing the history of the development of opto-mechanical scanners, it will describe a modern high performance thermal imaging system and consider some of the features which have contributed to its successful introduction into current projects.

PHYSICAL PRINCIPLES

All bodies at temperatures above absolute zero radiate electro-magnetic energy. Therefore in principle, if this energy can be collected and detected, a thermal image can be generated by re-arranging the signal from the detector to form a spatial analogue of the original scene. If the radiating object is a classical 'black body', that it, a perfect radiator, the spectral form of the emitted radiation as a function of temperature is described by Planck's equation as shown in Figure A.1 Appendix 1. Other bodies are less perfect radiators, thus a scaling factor has to be applied whose magnitude depends upon the physical nature of the object and its surface characteristics. This scaling factor, known as the object's emissivity, may itself be a function of wavelength. Objects which are non-perfect radiators must, by definition, be partially reflective and hence may contribute a reflected component to the radiated energy.

The wavelength of maximum emitted radiation of a black body is a function of its absolute temperature and is defined by Wien's Law as also shown in Figure A.1 Appendix 1. Other factors being equal, detection efficiency will be a maximum if the peak responsivity of the detector matches this wavelength of peak radiation. For the sun, at about $6000^{o}K$, the peak occurs in the visible band between 0.3 and 0.7 microns which, for evolutionists, provides one of the reasons for the natural selection of the response of the human eye. In the case of visual radiation, the sun's energy is selectively absorbed, reflected and scattered by different objects as a function of wavelength to create the phenonemon of colour vision in the human eye.

The peak of the radiation curve of a relatively hot object, such as a missile efflux, occurs in the band 1 - 5 microns whilst the peak radiation of objects near ambient temperature, such as a man, a tank or the natural terrain, occurs at about 10 microns.

Practical infra-red imaging systems designed to operate at or near these wavelength peaks of natural radiation have been called 'thermal imagers', IR Scanners or FLIRS (an acronym for Forward Looking Infra Red but widely and confusingly used to describe thermal imaging systems in general). In this paper we shall use the generic term 'thermal imagers' regardless of application.

The choice of operating waveband for thermal imagers is further constrained by the attenuation characteristics of the atmosphere. Examination of Figure A.3 Appendix 1 shows there are two 'windows' or regions of minimum attenuation, at 3 - 5 microns and at 8 - 13 microns. To the physicist, both these bands fall in the mid, or intermediate IR region; but by common usage, the thermal imaging community refer to 3 - 5 microns as the mid-IR band and the 8 - 13 microns as the far-IR band. This convention is adopted here.

Examination of the black body radiation curves in Figure A.1 Appendix 1 for hot and ambient temperature bodies shows that in absolute terms, a hot object actually radiates more energy in the 8 - 13 micron band than an ambient temperature object. Similarly it is apparent that an ambient temperature object also emits some energy in the 3 - 5 micron band. Nevertheless it is apparent that with equivalently efficient detectors, hot objects will be most efficiently detected in the 3 - 5 micron band whilst ambient temperature objects are most efficiently detected in the 8 - 13 band.

From the foregoing it will be evident that thermal imagers designed for applications which involve viewing the natural scene, or objects near ambient temperature, will be designed to operate in the 8 - 13 micron atmospheric window; whilst hot efflux tracking applications will more usually operate in the 3 - 5 micron band. There are exceptions to this general rule but these are generally limited to specialist applications or where a trade-off has been made between performance and the achievement of some practical advantage, e.g. elimination of cryogenic cooling to effect a cost/weight reduction, which influences the choice of detectors and waveband.

The key equations underlying the design of thermal imagers are included Appendix 1 and a more complete treatment can be found in Hudson (1) and Lloyd (2).

OPERATING PRINCIPLES

In considering how a thermal image is produced it is first important to recognise that any system will have two figures of merit; its spatial resolution and its thermal sensitivity. Whilst it will become evident that in fact these two parameters are inter-related, as an initial simplification they can be considered as 'how small an object can be resolved' and 'how small a temperature difference can be perceived'.

The search for thermal band imaging systems has been in progress for over a century. Whilst opto-mechanical scanning arrangements have been known for many years, the lack of adequately sensitive detector elements have prevented the achievement of militarily useful levels of imaging performance until the last decade.

There has been one notable exception to exclusive reliance on opto-mechanical scanning and this was based on the pyro-electric effect in triglycine sulphate. In a practical form the TGS target is mounted in a vidicon envelope fitted with a germanium window and the rear surface of the target is scanned by a conventional vidicon electron gun, as shown in Figure 1. By opto-mechanical scanner standards, pyro-electric cameras are relatively insensitive and of low resolution.

Although equipments are now available in a fully developed form, their limited performance precludes their use in most military roles and they are unlikely to find application for avionic purposes. All high performance thermal imagers are currently of the opto-mechanically scanned type. With these, one or more sensitive infra-red detectors are arranged on the focal plane of an optical system which itself comprises elements having optical 'power' and some form of optical scanning components. See Figure 2. The powered elements collect and focus the incoming radiation, whilst the scanning components provide a means whereby the thermal image of the object can be moved across the detector, or detectors, in a prescribed scan pattern. Subsequent electronic units amplify and process the signal current to reconstruct a picture for display, usually on a CRT for avionic applications. The visual image is almost invariably presented in a conventional T.V. format and the variation in intensity in the image is directly related to variation in radiation in the thermal scene. Whilst the basic concept of such a system is straight forward, the physical realization may be relatively complex and the forms of the practical embodiment are many and varied.

In the earliest designs, the whole image field was scanned across a single detector but later, multiple detector arrays evolved with, at the other extreme, 400 elements arranged so that each detector scanned only one line as shown in Figure 3. The diversity of scanning mechanisms has been even greater with mirrors, mirror polygons (both internal and external) and refractive elements being oscillated or rotated around almost every conceivable set of axes to effect the required motions of the image across the detectors. Subsequently, informed opinion became polarized into two competing camps, the first advocating so called parallel scanners (Figure 3) and the second, serial scanners (Figure 4). Current opinion has tended to settle on a compromise between these two extremes, known as the serial/parallel scanner (Figure 5) the efficacy of which has been reinforced by the latest developments in detector technology. Developed equipment based on parallel, serial and serial/parallel detector configurations are, or have been, used in avionic applications and so each of the arrangements will be briefly described.

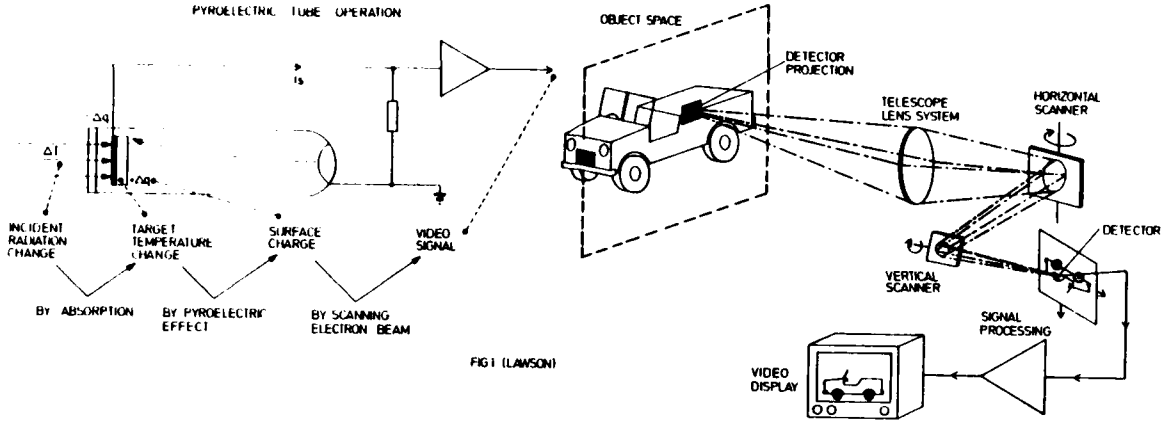


FIG 1 (LAWSON)

FIG 2

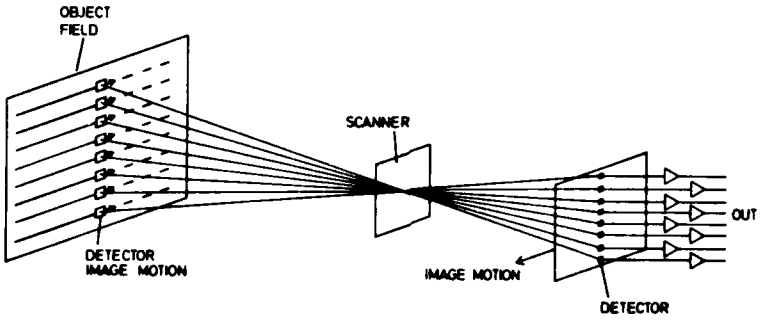


FIG 3

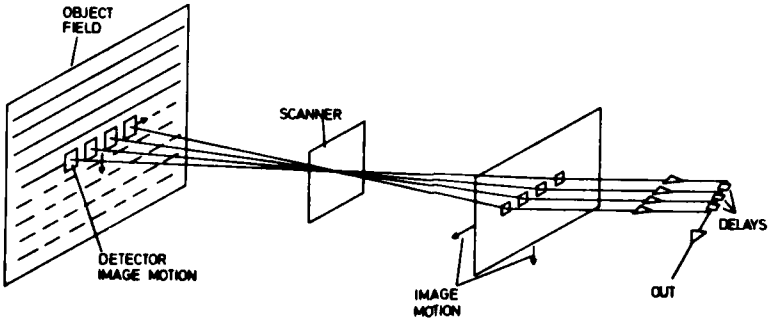


FIG 4

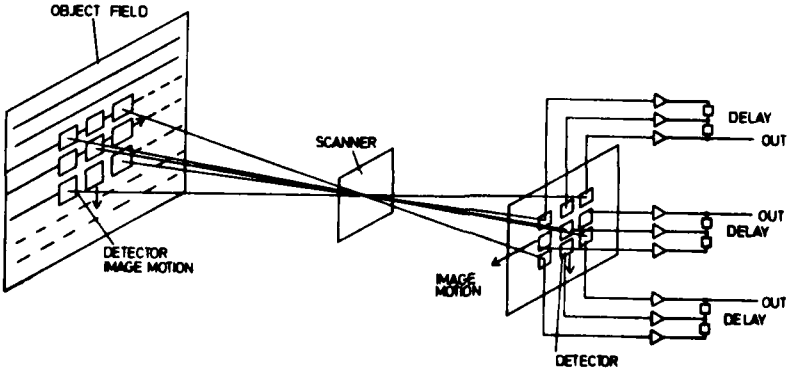


FIG 5

With the parallel scanner, the image is generated by a horizontal motion of the scene across a large vertical array of detectors. In this case the scan mechanism will operate relatively slowly for a given number of picture points per second in the displayed image. (Because, in effect, the information rate is the product of the number of elements and the scan rate). Virtually all thermal imagers designed for avionic use display a picture in a standard television format where a pair of interlaced fields constitute a frame of picture information. In some parallel scanners, a mechanism is included to displace each successive scan in the vertical direction by a half line pitch and thus to introduce a true interlace in the scanner itself. Other types produce a pseudo interlace, either by repeating the same information on odd or even fields, or by scanning the full frame of information, storing it and reading out the separate fields in sequence.

Because of the low scan rate, signal bandwidths in parallel scanners are low, reducing the required noise bandwidth and yielding a high signal/noise ratio. However, each detector requires a separate head amplifier and signal channel so that the advantage of a simple low speed scanner is offset by considerable electronic replication.

The serial scanner represents the other extreme in which a single detector, or a serial array of detectors, are scanned over every line in the field in turn. This arrangement involves exceptionally high scanning speeds, typically up to 160.000 r.p.m. and consequently demands very wide bandwidths. Indeed, the required bandwidths are so high that detectors are required to work up to, or even beyond their nominal cut-off frequency and, with such a wide bandwidth, signal/noise ratios are low. To combat the signal/noise loss it is usually necessary to employ a serial row of detectors aligned along the direction of scan and to combine their outputs by a process known as 'time delay and integrate' (T.D.I.) In effect, each signal is amplified, delayed by an amount proportional to scan speed and inversely proportional to its position in the row, and then added together. In this way the signals add arithmetically whilst the noise adds in an R.M.S. fashion to yield an improvement in signal/noise ratio proportional to $\sqrt{n}$, where n is the number of serial detectors.

Although only one electronic channel is needed after the T.D.I. the exceptionally high scanning speed presents special design difficulties especially in the severe environment of an aircraft. For these reasons, although compact serial scanners have been built and have found limited application, they are not the currently preferred solution for most users.

The solution most generally preferred today is the serial/parallel scanner which effectively combines some of the best features of the alternatives whilst avoiding the most extreme disadvantages of both. In a typical modern design 6 serial elements are arranged in 8 rows which are scanned across the picture to produce a swathe of information and down the picture to produce a number of contiguous swathes. At one time it would have been argued that such an arrangement would have produced unacceptable cosmetic defects (banding effects) in the image due to channel to channel non-uniformity. However, as we shall see later, modern signal processing techniques effectively eliminate such problems and a degree of signal uniformity is obtained which was previously only possible from serial scanners. In the serial/parallel scanner, the scan rates are relatively high, but within practically achievable limits, and the consequential high bandwidth falls well below the detector cut-off frequency. Furthermore the serial element, when combined using a T.D.I. function, improves the signal noise ratio by $\sqrt{6}$ to help compensate for the large noise bandwidth. In effect the serial/parallel scanner represents a balanced compromise with the total number of detectors chosen to effect a desired standard of performance.

An elementary analysis, presented in Appendix 1, shows that the performance of a scanner is essentially the product of the speed of the optics (f/no) and the square root of the total number of detectors used (neglecting scanner efficiencies and other losses). This 'performance' of the system can be distributed between the number of picture points in the field, thus influencing the spatial resolution, and the temperature sensitivity, since the product of the square root of the number of picture points times the sensitivity is a constant. To a first order therefore, and at a fixed f/no. and for a prescribed number of picture points in the image, the resultant sensitivity is independent of the scan format, and is only dependent on the number of detector elements used. Thus ignoring the effect of $1/f$ noise, which can be a problem at slow scan speeds, there should be little to choose in the performance achievable between parallel and serial/parallel scanners having the same number of detectors. This argument lends support to the contention, presented earlier, that the serial/parallel scanner solution enables a balanced design compromise to be reached without prejudice to performance. It was this reasoning that led to the basic design parameters chosen for the U.K. Thermal Imaging Common Module programme.

THE 'SPRITE' DETECTOR

In the discussion so far we have referred only to arrays of discrete infra red detectors. In 1981 a new form of detector, developed at Royal Signals and Radar Establishment (Malvern) and originally named the Tom Elliott Device (T.E.D) after its inventor Dr. C T. Elliott, was publicly announced (Ref 3). This device is now more generally known as the Signal Processing in the Element (SPRITE) detector. The advent of the SPRITE makes serial/parallel scanners undisputably the preferred choice since in effect it enables the whole serial array of discrete elements to be replaced by a single detector requiring only one electronic channel without the need for a separate T.D.I. function.

As illustrated in Figure 6, the detector takes the form of a strip of Cadmium Mercury Telluride (CMT) in which detected energy is swept along in the form of charge carriers at a velocity which matches the scanning speed. In this way the photon generated charge is delayed and integrated in the bulk material to yield an output signal current equivalent to a serial array of discrete elements after T.D.I. With this single invention, an imaging performance which would previously have needed 50 - 100 discrete elements could be achieved with a parallel array of only 8 SPRITE detectors. With this came a corresponding reduction in electronic complexity. The number of amplifier channels is drastically reduced and the T.D.I. circuits are eliminated altogether. This important advance has had the most significant implications on the cost, weight, size and complexity of the latest thermal imager designs.

It is interesting to note that the new detector has removed one degree of freedom from the sensitivity equation. As the SPRITE's ability to integrate is determined by the physical constants of CMT, the integration period is fixed in time. Thus the higher the scan speed chosen, the greater the length of the device contributing to the integration of the signal. Applying this fact to the sensitivity equation it can be shown to a first order of accuracy, that the sensitivity of an imager is independent of the number of SPRITE detector elements, providing the scan speed is increased proportionally to keep the information rate constant. See Appendix 1.

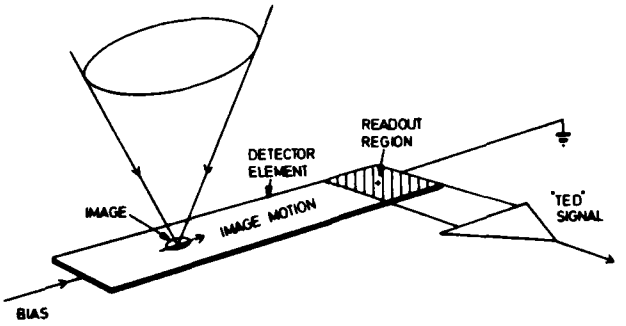


FIG. 6

A MODULAR APPROACH TO THERMAL IMAGERS.

There has been a growing realization, in both the USA and Europe that special to type development of thermal imagers for each and every application was both impractical and uneconomic for all but the most specialist applications. This has led to the philosophy of a modular design approach whereby sets of equipment building blocks have been designed which could be configured into complete systems to suit the special needs of individual projects. Not only has this approach reduced the non-recurring cost and timescale of new projects, but it has also enabled the cost savings of production scale to be realised as well as bringing substantial benefits in both logistics and product support. Furthermore, the modular approach provides the opportunity to take advantage of technological developments during the life of the equipment based on the substitution of replacement modules, thus extending the performance and the service life of the product.

The USA was the first country to embrace the modular imager concept as policy with D.O.D. approval being given to the project in 1974. The U.K. followed late in 1976 and France started her own programme in the same year. All three programmes followed the same general philosophy attempting to produce a set of basic building blocks from which systems contractors could configure customised equipments to meet specific applications. The original intention was to meet the requirements of all three services although the degree to which such standardization has been achieved varies from country to country.

In the USA, after considering the alternatives, the parallel scan approach was adopted. This used conventional C.M.T. detectors and a slow scan mechanism to achieve high performance; but at the expense of a large detector array (60 to 180 elements depending on the application). As a consequence a correspondingly large number of electronic channels are required and for an indirect view, T.V. compatible output signal, a full frame storage scan converter is needed. In its original form, a Vidicon camera was involved in the scan conversion process, but fully digital scan converters have since been produced. In Europe, initial attempts to establish a collaborative programme failed and France and the U.K. proceeded with national programmes based on rather similar serial/parallel scanning systems. Germany adopted the U.S. common module imager. In the U.K. the Thermal Imaging Common Module (TICM II) programme started by adopting a 48 element serial/parallel scanner but part way through the development programme prototypes of the SPRITE detector became available for evaluation and comparison. At this stage an element of risk was involved in adopting so new a product, but such were the advantages that the decision was made to abandon discrete detectors and base the design on the new detector. In the event the decision has been amply justified and the TICM II development was completed to plan embodying the SPRITE detector.

SIGNAL PROCESSING & CONDITIONING

In the thermal band a typical natural terrain scene may have a total dynamic range varying between 1 or 2°C to 10°C or even more on some occasions. There may be hot targets within this scene extending the range by 10 or 20°C and sometimes much more. The sky may be in view which, on a cloudless day, can radiate temperatures well below 0°C. When compared with the thermal sensitivity of the imager, which may be typically 0.1°C, it is evident that the total dynamic range within the field of view may be very large indeed. See Figure 7. When compared to conventional television cameras which can usually encompass the whole range of natural scene brightness, reproducing the absence of light as black and bright parts of the scene at peak white, it is evident that thermal imagers need to process the video signal in a fundamentally different fashion. For example, it is often not desirable to attempt to reproduce the whole dynamic range in the thermal scene. Only the part of it embracing the subject of interest is usually displayed. This has led to evolution of the concept of a thermal window (Gain) and a temperature offset (Offset). The former sets the range of temperatures that will be displayed and the later the temperature that will be reproduced as black in the displayed image. This may be absolute or relative to the scanner body temperature depending upon how it is implemented. Thus it is possible to select any temperature 'window' from the available dynamic range and display that as a scale of grey on a black and white monitor. It should be noted that, within a prescribed thermal band, there are no sharp spectral variations and so concepts of colour, analogous to the visual band, have little meaning. Thermal imaging is, to all intents and purposes, a monochromatic process and therefore it is perhaps fitting that it should be reproduced on a black and white display.

In the 8 - 13 micron band during daytime, reflected sunlight makes little contribution to the received energy. Solar heating on the other hand, can significantly affect the picture contrast and the cold sky, reflected off low emissivity surfaces, frequently makes a large contribution to contrast. In the 3 - 5 micron band reflected solar energy is significa.. and adds to the complexity of interpreting the image.

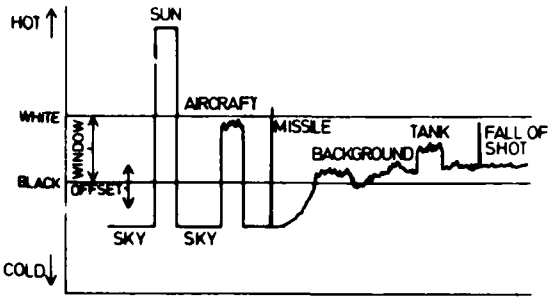


FIG 7

Despite the comments made in this paper emphasising the difference between visual and thermal band scenes, it is little short of remarkable that on most occasions the general appearance of thermal pictures are subjectively so similar to a television picture.

To the operator the thermal band brings not only an independence of light levels but many incidential advantages. Militarily, perhaps the most obvious example is that of conventional camouflage, which may be perfect in the visual band but almost completely ineffective in the thermal band. Most smokes and all but the most dense fogs are relatively transparent in the thermal band allowing viewing ranges to increase by a factor of 3 in moderate fog and by 10 in mist.

Another factor of significance in the design of the signal circuits, which arises out of the nature of the thermal scene, is the 'DC' pedestal in the detector signal current and the 'AC' component arising out of differences within the scene. Not unusually the 'AC' signal may be only 1% of the DC component or less. The DC component effectively represents the absolute radiation from the ambient background, whilst the AC component represents the variation in radiation within the scene, which is eventually to be displayed as a black and white picture. Equally, the variation in DC pedestal and AC gain (responsivity) between one element and another, has always presented problems in imager design. Essentially the requirement is to dynamically match every detector channel so that no perceptable variation is evident on an otherwise uniform scene and to present the AC signal on the display, so that a pre-determined level of thermal radiation corresponds exactly to a prescribed grey scale value independent of gross variations in scene content. The satisfactory achievement of this requirement represents one of the more significant advances in thermal imaging technolgoy.

This area is one in which the terminology has been rather inexact, with the terms 'matching', 'referencing' and 'clamping' being used virtually interchangeably in general usage. In the U.K. there is now some concensus to use 'clamping' to mean the electronic process of restoring the DC component to an otherwise AC coupled signal.Clamping will normally be carried out, at some stage after signal amplificatio , when all detectors are seeing the same input signal level, which may or may not be a defined absc.ute level. Referencing on the other hand is the process of relating the absolute level of the sign.l on all channels to a pre-determined datum. This datum may be the temperature, controlled or otherwise, of an object which the detector 'sees' during the clamp period, or for example, it may be an electronically computed value equal to the mean signal, derived from the scene being viewed (which is similar to AC coupling). Yet again it may be another reference chosen by the systems designer (e.g. an area in the centre of the image). Matching is the process of reducing the residual differences between channels after the clamping and referencing processes have been accomplished.

All three of these techniques must be implemented successfully to achieve a consistently uniform picture. Furthermore for a 'common module' concept to be effective, the final choice of clamping and referencing means must be left to the systems designer to enable him to configure the circuits to suit the application.

With the exception of the introduction of the SPRITE detector the successful reduction of inter-line and interband (swathe) non-uniformity has been more important to image quality than improvements in resolution and sensitivity in the last few years. The Class II U.K. Common Module Programme has taken this process a stage further with the full automation of the channel matching and the window selection process; a combination believed to be unique to that programme. This development has far reaching consequences both for the achievement of full modular interchangeability, but more importantly in avionic applications, for completely hands-off operation of the imager in operational use.

DESCRIPTION OF U.K. CLASS II TICM..

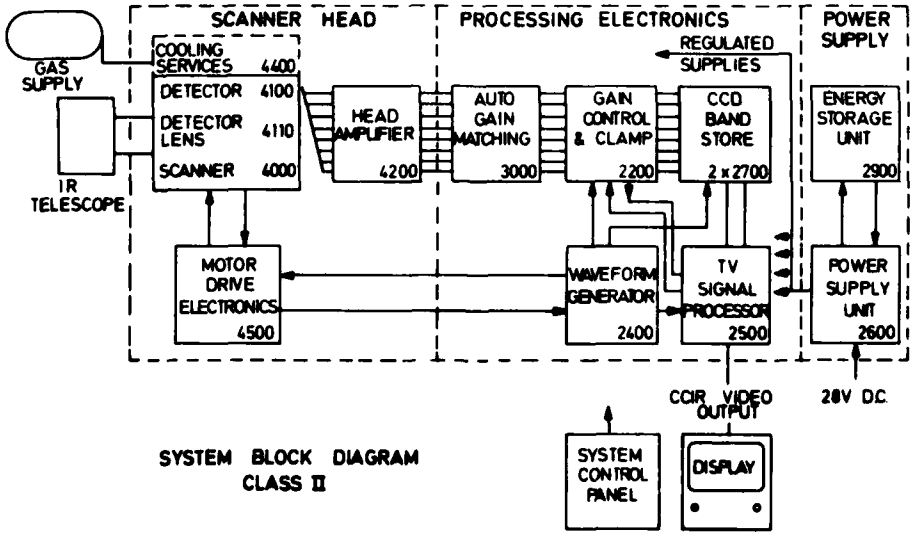
As an example of a fully developed state of the art design now in full scale production, the TICM II will now be described. Figure 8 shows a block diagram of a complete imager configured from the modules. Essentially it consists of 3 main functional blocks, a telescope, a Scanner Head and a Processing Electronic Unit with a Power supply. Controls and displays may either be physically integrated with the host vehicle or with the imager itself depending on the application.

The magnification of the telescope, which for use in the 8 - 13 micron band uses germanium as the principle refracting material, is chosen to match the basic 60^{o} Scanner field of view to the field of view (F.O.V) required in the application. Thus a X3 telescope matches the scanner to a requirement needing a 20^{o} F.O.V. Currently single or dual field of view optics are the most common choice. Quite frequently with dual magnification designs the angles are chosen so that a target first detected at the wider angle may be recognised on the narrower angle setting. Zoom optics, and a simpler variant which zooms between 2 pre-set positions (but is not controlled at intermediate focal lengths) are available, but so far their cost militates against widespread usage.

The scanner is the opto-mechanical heart of the system, with its rotating polygon and oscillating mirror and associated drive electronics generating the relative motion between the detector and the scene. In TICM the mechanisms for generating the scan motions along orthogonal axes are quite independent, to facilitate subsequent enhancements; but more importantly, in the short term, enabling different scan arrangements to be chosen for different applications. For U.K. and European use a 625 line 50Hz field compatible scan is employed whilst for U.S. applications the relative scan speeds are adjusted to suit the 525 line 60Hz standard. The scanner head also contains the detector and 8 associated head amplifier channels.

In common with all high performance thermal imagers the detector is cooled to cryogenic temperatures ($77^{o}K$) to minimise the number of thermally generated carriers and thus maximise the signal/noise ratio for a given input flux. Cooling is commonly achieved by using a Joule Thompson mini-cooler supplied from either a bottle containing gas under pressure or a small local compressor. The cooler works on the principle that gas under pressure cools as it expands through an orifice. The cold gas also cools the incoming gas via an integral heat exchanger and this process continues in an iterative fashion until upon expansion the gas finally liquifies. The cryogen so produced then cools the detector.

The present cooling arrangements, whilst entirely satisfactory in performance terms, do incur the logistic disadvantages of gas supplies and the mass and bulk penalty of carrying bottles. Compressors although practical in some static installations are heavy and need servicing, as does the associated gas cleaning plant. For the future, self-contained closed cycle cooling engines based upon one of several thermodynamic cycles will undoubtedly enable the present arrangements to be replaced. The engineering design of these engines, which present formidable problems in achieving reliable, vibration free operation, is proceeding in a number of countries and their advantages are such that a practical design can only be a matter of time.



SYSTEM BLOCK DIAGRAM
CLASS II

FIG 8

Consider now the signal processing chain which is illustrated in Figure 9. The input signals from the detector are AC coupled to avoid the DC offset associated with the signal current pedestal previously described. The head amplifiers provide sufficient gain to raise the signal to a level that swamps subsequent noise sources. The channels are then automatically equalised to remove gain and detector responsivity variations, the signal is clamped to re-establish the DC level discarded by AC coupling and then referencing and matching is implemented as has been discussed. With the DC levels re-established, and an accurate reference introduced, the signals can now be windowed to select that portion of the total signal excursion required for display. Up until this stage the signals are produced in a frame of reference determined by the IR scanning parameters and synchronised accordingly. The signals have next to be transformed by a parallel to serial converter into a serial information stream which is timed in accordance with the exacting phase stability requirements of a conventional television signal. Finally, synchronising signals are added to produce a composite television signal which is fully compatible with T.V. displays and recording equipment of the appropriate line standard.

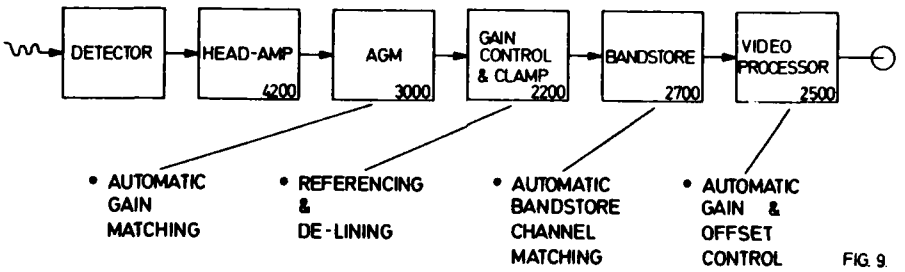


FIG. 9

Having presented the overall system some of the more important features can be identified for comment.

Despite stringent process control CMT detectors exhibit variation in responsivity and spectral response, both between individual elements in an array and, of significance to interchangeability, between arrays. Amplifiers too exhibit gain variation between channels due to component tolerances, over the military temperature range, and with life. Unless all channels are matched to an accuracy substantially less than 1% throughout their life, cosmetic defects are visible in the displayed image in the form of a repetitive line structure. In earlier designs a manual gain setting process has been employed but this is costly in terms of test time, the need for on-condition maintenance and because it prevents direct modular interchangeability. In TICM II the function is fully automated in the Automatic Gain Matching module, operating off the available video signal, to eliminate the need for manual adjustments. Having taken great care to eliminate variation in the amplifiers and in the subsequent clamping and referencing process, residual differences in the signal still remain principally because of the spectral response variations between detectors. This has necessitated a further stage of automatic compensation or de-lining, which is carried out in the Gain Control and Clamp module.

With the 8 parallel channels carefully matched, the parallel to serial converter presents its own sources of gain variability, particularly those concerned with the bandstore effecting the scan conversion process. There again the channel balancing process has been fully automated with individual channel gains being set, under microprocessor control, as the equipment is being switched on and thereafter monitored continuously during operation. As a consequence the high degree of channel matching established in the preceding modules is maintained through the bandstore modules. From this point onward the signals are in a serial form and relatively tolerant to circuit variability.

The final module of interest is the T.V. signal processor which as might be expected, performs the usual functions of blanking insertion and sync mixing. In addition, recognising the system designers' needs to incorporate symbology and alpha-numeric data on the displayed image, it also provides signal mixing facilities enabling these external inputs to be inlaid or overlaid on the video. However. more importantly, it is in this module that the selection of that part of the thermal scene which is to be displayed can be established. Selection of the thermal window (Gain) and its absolute level (Offset) can be controlled manually; but a fully automatic control mechanism is also incorporated. For avionic use the automatic feature is quite essential. The airborne scene can change with great rapidity, for example, presenting a view of the cold sky at one moment and a changing terrain view at another. Under these conditions, considerations of pilot workload preclude normal manipulation of the Gain and Offset controls and, even if he had the time, under these dynamic conditions it is unlikely that manual adjustment would be consistently optimal. The need for automated operation has been recognised for years, but only recently the availability of micro-processor control systems has permitted the provision of automatic control as an integral function. In TICM II control of this facility is resident in the T.V. Signal Processor module, where the processing algorithms are embedded in software and where selection of the area in the viewed scene which is to be used to generate the control signals is made. In keeping with the module philosophy, the system designer has the freedom to specify software algorithms and control areas to optimise the design to meet his application needs.

To complete the design of a custom built thermal imaging system, only housings, a control panel and a display are needed.

CONCLUDING COMMENTS

The principal airborne applications of thermal imaging equipment are either as a flying aid or for target acquisition and weapon aiming. Most usually the flying aid role is fulfilled by a fixed, forward looking imager, with a field of view matching the Head Up Display. With the modern holographic HUD, horizontal angles of up to 30° are typical with the thermal image displayed so as to overlay the real world in one to one coincidence when seen through the HUD combiner. The same wide angle system may be used for target acquisition either visually or automatically by extracting target information from the video signal.

The recognition and tracking of small targets invariably requires the use of a narrow field of view, with sightline stabilization. It must usually be steerable over a solid angle at least as great as the acquisition field. Installations may either be within the airframe or podded as suits the application.

There are of course, many possible variations of these basic systems involving integration with other sub-systems such as laser rangefinder, auto trackers, auto recognisers and other visual aids. The combinations are limitless but in all cases thermal imaging adds a significant new dimension to the airborne environment. It increases utilization of the aircraft and enhances the weapon delivery capability.

As we have attempted to show in this paper, in the last few years there has been steady progress in the provision of equipment fully engineered for the airborne environment. With the advent of fully automatic modular equipment and the SPRITE detector thermal imaging can now be considered a mature technology.

References:

1. Richard D. HUDSON Jnr. - 'Infrared Systems Engineering'.
2. J.M. LLOYD - 'Thermal Imaging Systems'.
3. IEE Conference Publications No. 204. Advanced Infrared Detectors & Systems - October 1981.

Acknowledgements:

The author wishes to thank the many sources of information used throughout this review, part of the work having been supported by Procurement Executive, MOD.

APPENDIX 1

The radiation emitted from a black body is given by Planck's Equation.

$$W_\lambda = \frac{c_1}{\lambda^5} \frac{1}{\exp\left(\frac{c_2}{\lambda T}\right) - 1} \quad W\,cm^{-2}\mu m^{-1}$$

W_λ = spectral radiant emittance

T = absolute temperature of the body

$c_1 = 2\pi h c^2$ where h = Planck's constant

$c_2 = c\,h/k$ c = Velocity of light

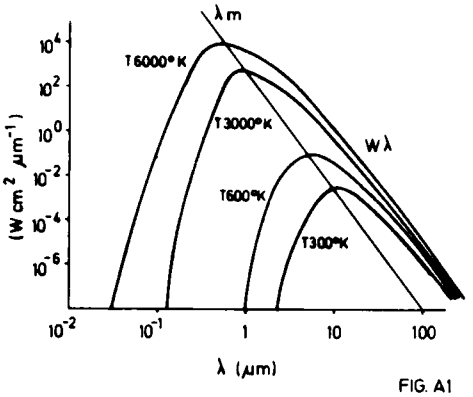


FIG. A1

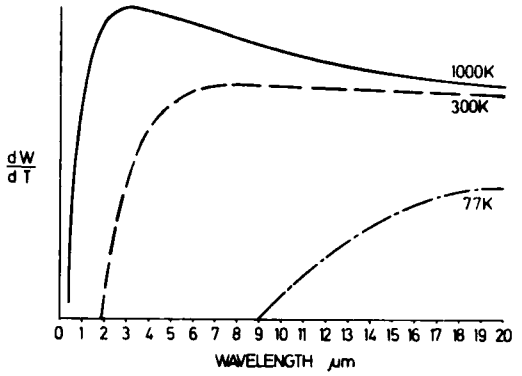


FIG. A2

Integrating over all wavelengths gives the radiant emittance, the flux emitted per unit area

$$W = \frac{2\pi^5 k^4}{15 c^2 h^3} T^4 = \sigma T^4 \quad W\,cm^{-2}$$

σ = Stefan-Boltzmann constant

Differentiating Planck's Law and solving for the maximum gives the wavelength of peak emission (λ_m) for a black body of absolute temperature T, Wien's Law.

$$\lambda_m = 2898 / T \quad \mu m$$

Taking a partial derivative of Planck's Law with respect to temperature gives the differential radiance, a measure of contrast. This can be solved numerically but for $\lambda T < 3100$ $\mu m\,K$ is given to a good approximation by

$$\frac{\partial W_\lambda}{\partial T} = W_\lambda \frac{c_2}{\lambda T^2}$$

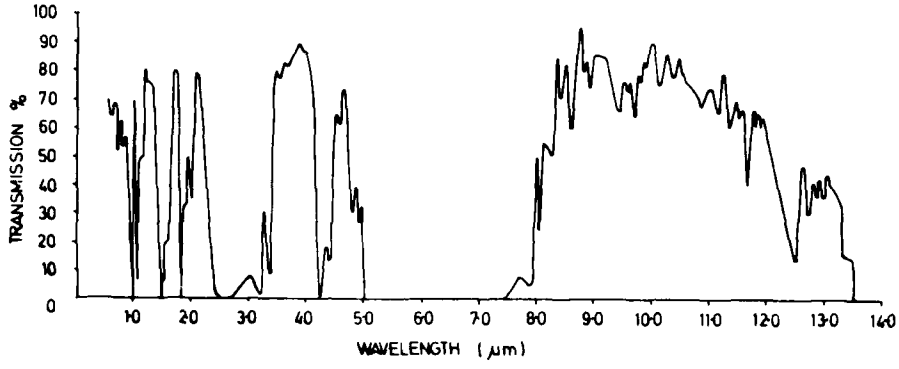
Solving these equations for a black body at 300°K, in the 8-13μmband (changing units).

Total radiant emittance = 150 $W m^{-2}$

Differential radiant emittance

for 1°C temperature range = 0.1 $W m^{-2}$

A typical atmospheric transmission curve is shown in Figure A3.



TRANSMISSION OVER ONE SEA MILE. FIG A3

Photon detectors have a theoretical response curve shown in Figure A4.1. More typically a real detector is combined with the transmission of the dewar window and will look like Figure A4.2.

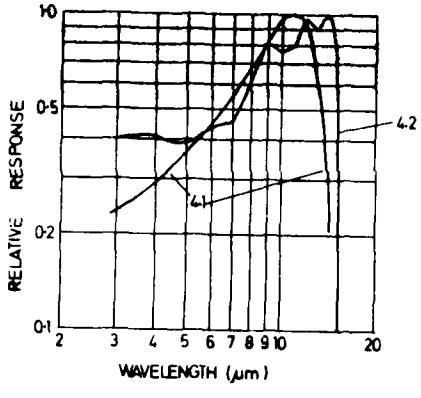


FIG A4

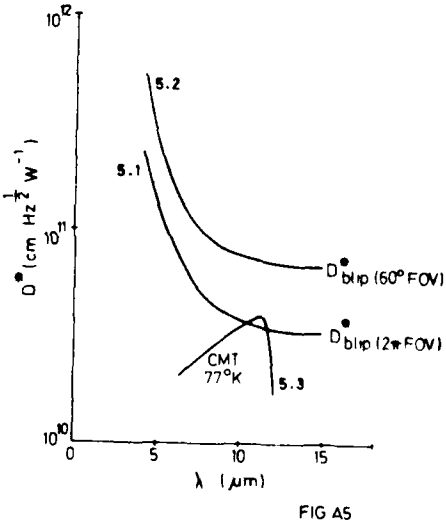


FIG A5

The ultimate performance limit of these detectors is defined by fluctuations in the radiation from the ambient background, - the Background Limited Infrared Performance (BLIP) Figure A5.1. The absolute level of this BLIP can obviously be controlled by limiting the field of view of the detector to that required by the system. Provided the 'shield' which limits the detector's field of view is cooled to cryogenic temperatures the total background flux incident on the detector can be reduced and the noise fluctuations will correspondingly be reduced. Figure A5.2. Realisable detectors can approach the BLIP limit Figure A5.3.

System performance evolves logically from detector performance.
The Noise Equivalent Power of a detector, NEP, is the signal required to give a SNR of 1:1

$$NEP = H\,A\,\frac{Vn}{Vs}$$

H = Irradiance W cm^{-2}

A = detector area

Vn = RMS noise voltage

Vs = RMS signal voltage (fundamental)

Inverting to give a figure of merit which is greater the better the detector, gives Detectivity

$$D = \frac{1}{NEP}$$

Normalising for root area, to which Detectivity is inversely proportional in a photon detector, and root bandwidth gives a figure of merit which enables comparisons to be made between different detectors in different systems, D* (pronounced D - STAR)

$$D^* = \frac{(A\,B)^{\frac{1}{2}}}{NEP} \qquad cm\ Hz^{\frac{1}{2}}\ W^{-1}$$

A = detector area B = Bandwidth

D* is traditionally quoted at peak sensitivity $D^*_{\lambda pk}$ or as an integrated figure against a target at 500°K, D^*_{500K} and at a specified temporal frequency e.g. D* (500K, 20kHz).

Occasionally this may be further extended to include an atmospheric transmission dependency and the expected target temperature. This is called M* (or M - STAR).

$$M^* = \int_0^\infty \left(\frac{\partial W_\lambda}{\partial T}\right) ta_\lambda\, D^*_\lambda\, d\lambda$$

ta_λ = atmospheric transmission

However, M* is less frequently used than it used to be, analysis being more frequently undertaken from the more fundamental parameters.

Adding the photon gathering power of the imager optics, the efficiency of the scanning mechanism and, where appropriate, the atmospheric transmission, gives a system figure of merit.

The traditional figure of merit is the Noise Equivalent Temperature Difference NETD.

$$NETD = \frac{4F^2 B^{\frac{1}{2}}}{A^{\frac{1}{2}} \int_0^\infty \left(\frac{\partial W_\lambda}{\partial T}\right) to_\lambda\, ta_\lambda D^*_\lambda\, d\lambda}$$

to_λ = Optics Transmission

F = System F/No

If atmospheric transmission is neglected then NETD becomes 'zero range NETD'

note $B \propto \frac{1}{N}$

N the number of parallel elements in the detector $\therefore$ NETD $\propto \frac{1}{\sqrt{N}}$

which in the form shown is hidden in the Bandwidth factor. (B)

Also where TED devices are employed, to a first order the D* increases with root scan velocity exactly offsetting the reduction in numbers of elements required to scan a fixed number of picture elements in a given time. Thus the $1/\sqrt{N}$ no longer applies, providing the scan velocity is increased inversely proportionally to the number of TED detectors.

More recently, system performance has been specified in terms of the ability of an observer to 'just observe' a given test target. Two forms of target are used, a circle and a fixed aspect ratio 4 bar target. The former when plotted against spatial frequency is known as Minimum Detectable Temperature Difference, MDTD, the latter the Minimum Resolvable Temperature Difference.

ADVANCES IN AIRBORNE RADAR - THE NEW CAPABILITIES

by
C.M. Stewart
Chief Engineer
Radar Systems Department
Ferranti plc
Edinburgh

SUMMARY

The power and speed of the minicomputer, microprocessor, and other signal processing sub-systems have had a very significant effect on the operational capability now available from airborne radar. The mechanisation of hitherto unrealisable processing strategies has encouraged the development of more sophisticated pulse compression and pulse doppler radars and allowed them to achieve their full potential within the size and weight constraints of an aircraft installation. The paper reviews the new capabilities in each of the operational roles in which radar is used outlining the techniques employed.

1. Introduction

Despite the development of alternative technologies, radar has retained its attraction as a sensor because of its flexibility in adapting to a wide range of operational roles, its relative immunity to weather conditions and its unequalled ability to detect relatively small targets at considerable distances and within a wide field of view.

Radar theory, although a relatively mature field of study, is still evolving, notably in the fields of detection in sea clutter and the application of newer forms of pulse doppler to air-to-air detection but like other branches of avionics the real driving force for recent developments has been the advent of powerful, high speed computing capacity with a size and weight suitable for airborne use. New radar capabilities, already predicted by theory, are therefore now realisable within the difficult constraints imposed by an aircraft installation.

A related development is the increasing adoption of coherent radar systems relying on the use of an R.F. power amplifying tube rather than the self oscillating magnetron. The new computing power available enables the capabilities of this type of radar to be fully exploited particularly in detecting weak target returns in a high clutter environment such as arises in the AEW role or air-to-air interception over land.

In parallel with these strictly radar developments has been the drive to obtain the maximum benefit from the variety of sensors carried on modern aircraft by integrating their outputs and controls at a common work station. This has led to the development of the digital bus as a highway for communication and between sensors including radar and the consequent need to standardise the interface structure of each sensor.

Attempts have been made to standardise the wide band video links between the sensors' video outputs and a common display or displays, permitting interchangeability between sensors. A raster scan T.V. format is usual for these links but in the case of radar it is necessary to interpose a scan converter between the raw video output and the video link and this is frequently combined with a digital store within which integration of successive p.r.f.s and scans takes place. This store replaces the long-persistence phosphor C.R.T. on which integration took place in earlier radar systems. More sophisticated radars, particularly of the maritime surveillance and interceptor types, go further and transmit to the display only the presence or otherwise of the target and its position on the display, the operator no longer being required to act as a detector but being replaced by automatic detection circuits within the radar processor. The design and proving of the algorithms setting the thresholds within these circuits is an important part of radar development since they must provide the maximum sensitivity with the lowest false alarm rate.

Although some of the new aspects of technology development already mentioned are common to all applications of airborne radar this is not always so and it is therefore convenient to continue this survey in terms of the separate operational roles of radar, tracing their growth in capabilities where this has occurred and attempting to sketch the technology which has permitted it to happen.

2. Air-to-Ground

2.1. Roles

The principal Air-to-Ground roles in which radar is involved are:

(a) Strike and Interdiction

(b) Reconnaissance

(c) Terrain Clearance and Close Air Support

However, it is in these roles that electro-optics competes with radar as a sensor and various forms of T.V., lowlight T.V., thermal imagers and laser systems have been developed in recent years to provide the high resolution and accuracy essential to these applications. Despite this, radar has retained an important position in this area for a number of reasons, amongst which probably the most important are:

(1) Only radar can provide in a single sensor both air-to-air search so that the aircraft can defend itself, and a useful air-to-surface capability for attack or navigation.

(2) With the right choice of frequency, radar can provide all-weather operation and is unaffected by battlefield smoke and fog.

Although some effort has been made to combine microwave and electro-optical sensors and obtain the best characteristics from both, this has so far not been very successful, principally becuse they compete for the same forward looking aperture on the aircraft. Also, because they use different parts of the frequency spectrum, no material for use as a "window", satisfactory to both, has yet been developed. Where the use of an electro-optical sensor is essential, current trends are to pod mount it beneath the wing or fuselage.

2.2. New Forms of Transmitted Signal

Until recent years almost all air-to-surface radars were non-coherent (with the possible exception of sideways looking Radars), i.e. they made no attempt to relate the phase of the transmitted and received signals. The significant changes of late are the moves firstly to pulse compression systems - representing probably the ultimate form of non-coherent system, and more recently to coherent systems enabling completely new capabilities to be developed.

Pulse compression provides a method of obtaining a short pulse and high mean power, without incurring the problems of handling high peak power. This short pulse gives the high range resolution essential in air-to-surface modes and figures of the order of 5 metres are not uncommon (equivalent to a pulse length of 30nS). Such systems can no longer use the magnetron as the main R.F. power generator and it is replaced by a stable low power oscillator driving a high power R.F. amplifier - currently a Travelling Wave Tube (TWT) would most likely be employed. A useful by-product of this change is that the frequency of transmission can now be changed very rapidly, on a pulse to pulse basis if desired, over a region within 5-10% of the nominal frequency. This provides a useful ECCM feature or can be used to minimise interference from other radars and R.F. sources. The low peak power distributed over a fairly wide frequency band makes this form of transmission more difficult to detect and constitutes the basis for the characteristics of Low Probability of Intercept (LPI) radars.

One disadvantage of the Pulse Compression System which arises from the long transmitted pulse is the large value of minimum range which is an inherent consequence of the technique. Since this minimum range could be of the order of 3000 metres, which could be inconveniently large at the intermediate range scales, non compressed short pulses are sometimes interleaved with the long pulses to fill the gap.

2.3. Improving Angular Resolution

Although, therefore, means are available for obtaining high range resolution, comparable angular resolution is more difficult to achieve since the size of antenna which can be accommodated in an aircraft is severely limited. Such resolution is however clearly essential if ground features and ultimately targets are to be recognised. In the past the only real solution to this problem was to use higher frequencies J band (~15 GHz) or even Q band (~35 GHz) instead of the usual X band (~9 GHz), despite the disadvantages of poorer weather penetration.

A method developed with the advent of monopulse radars was Monopulse Resolution Improvement (MRI) or enhancement which, provided a small enough range cell was used to separate targets in range, could locate the position of a point target within the fairly wide beamwidth of the antenna. (See Fig 1). Targets not on or near the boresight could then be rejected or in more sophisticated systems painted in the same position on the display with each successive transmitted pulse as the beam swept across them. More modern versions of this system use a digital frame store to hold the returns instead of painting them directly on the display.

"Monopulse Resolution Improvement" in fact does not provide an increase in time resolution in the sense that multiple targets in the same range resolution cell appear as a single response. However the ability to discriminate isolated targets is greatly improved. A genuine improvement in angular resolution may be obtained by the use of Doppler Beam Sharpening (DBS).

DBS operation requires coherency of the transmitted waveform which may be obtained either by amplifying a stable oscillation or by measuring the phase at the instant of transmission and correcting the echo phase appropriately. The effective beamwidth can readily be reduced by a factor in excess of 30 to provide tangential resolution comparable with the radial resolution of a few metres obtainable by pulse compression. These resolutions may be obtained at typical stand-off ranges of 20 to 40 km.

The techniques (See Fig 2) rely on the principle that when the antenna is pointed at an angle to the aircraft track, all ground returns within the beam but at the same range have different velocities relative to the aircraft. Spectrum analysis of such a range cell will show that the doppler frequency spread of ground returns from the front to the rear of the beam might typically be of the order of 750 Hz (centred on 10,000 Hz) for an aircraft flying at 300 Kts. A digital filter-bank may be economically implemented using the Fast Fourier Transform algorithm. Digital filters combine extreme stability with very flexible control of shape and impose no limit on the achievable resolution which is dependent on the aperture and the quality of data available for motion compensation.

The technique reverses ideas on some of the usual factors favouring high resolution. For example it is evident that the smaller the aperture used, the wider the beamwidth but the longer the target is within the beam. Since the resolution obtainable is a function of the time the target is illuminated, higher resolution is obtained with the wider beams generated by small apertures.

There are of course some constraints on such systems. Beam sharpening is not available on or near the aircraft track i.e. dead ahead, the resolution improvement being greatest with the antenna pointed at an angle of 90° to track.

Although the magnetron based radar is normally non-coherent, recently means have been found to achieve a degree of coherence by automatically adjusting the phase of the returning signal. This means that the advantages of DBS can now be obtained with magnetron radars and would allow many existing radars to be retrofitted with such a system.

DBS is usually applied in two ways; either as a means of improving the quality of an area ground map or in a spotlight mode where the ultimate angular and range resolution can be applied to a small area to permit recognition of target shapes. The value of such modes can be increased still further by using the velocity filters forming part of the DBS processing to identify moving targets on the ground. Such targets can then be surveyed by the spotlight mode to obtain further information.

2.4. Reconnaissance

Although the methods described earlier have application to reconnaissance, it has been more usual in this role to employ a full Synthetic Aperture Radar to generate a strip map of an area parallel to the path of the aircraft carrying it. This is done by either pointing the antenna orthogonally to the aircraft track or at an angle to the track (squint mode strip mapping) and building up a picture of the terrain in a series of consecutive strips, which are recorded on magnetic tape.

Considerable development of this technique has taken place over the last decade and imagery giving excellent optical correlation has been obtained.

The principle is similar to that already described, the range bracket of interest being divided into a large number of range cells each of which is analysed for its doppler spectrum. Successive spectra may be averaged in a ground stabilised store to reduce the effects of scintillation which become troublesome at high spatial resolution. Current practice is, of course, to perform these functions digitally. In the case where the antenna is pointed at right angles to the track, the doppler filter bank straddles zero frequency but for squint mapping, a specific doppler frequency is selected corresponding to a particular aircraft velocity and squint angle. Because the angular resolution obtainable by these sideways looking techniques corresponds to a much larger antenna aperture than is in fact used, the term "Synthetic Aperture Array" has been coined.

These systems require an accurate knowledge of aircraft accelerations in all axes to allow the application of corrections to the store addresses, thereby ensuring correspondence between assigned integrating store cells and individual areas on the ground for each transmitted pulse. Additionally and more critically the propagating path length change over the interval must be measured or calculated to allow correction of the echo phase. If such corrections are not made, smearing of the imagery will take place should the aircraft depart from a constant speed and direction due to atmospheric disturbance, manoeuvres etc. As before, the ultimate limit on angular resolution is the time available for integration which in turn is a function of cell size, antenna beamwidth and aircraft speed amongst other factors.

Since virtually all the processing described is performed digitally this has made it possible to apply new techniques for pattern recognition, the stored terrain image being scanned for targets of particular interest which can be described in terms of specific distributions of reflectors; convoys, tank formations and missile batteries are examples.

2.5. Terrain Clearance and Close Air Support

2.5.1. Terrain Clearance

The radar systems and fire power available to combat strike aircraft intrusions within enemy territory whether for local attacks at the edge of the battle zone in support of own troops, or deeper interdictive strikes, has led to the use of very low level approaches to targets, relying on radar to ensure that the flight profile remains within 60-200 m of the terrain profile as far as possible.

Although radar guidance techniques for achieving flight at low level vary somewhat between different systems, it is convenient to use the collective term "Terrain Clearance" to describe them, since all must at least fulfill this fundamental requirement.

The basic principle of all Terrain Clearance Systems is the same; to determine the profile of the ground ahead out to a distance at least compatible with the vertical manoeuvre capability of the aircraft and crew, compute the flight vector which will allow the aircraft to clear the terrain by the predetermined clearance height, and indicate this vertical flight direction to the pilot or autopilot. Where the desire is to hold the flight profile parallel to the ground profile but within certain preset g limits e.g. - 0.5g to +3g, the system is described as Terrain Following and these systems have probably had the most extensive development in recent years e.g. in the Tornado Strike Aircraft.

In such systems the usual procedure has been to perform a vertical scan over a limited angle in elevation and measure the range along the boresight to the interception point with the terrain. Since this interception point runs along the profile of the terrain as the antenna is scanned, the complete ground profile is obtained. It is necessary in such systems to ensure that the vertical scan is centred on the aircraft track to allow for drift angle and there may be restrictions on the aircraft turn rate because of the lack of adequate azimuth field of view.

Further developments of Terrain Clearance systems have attempted to remove some of the restrictions and in particular try to offer the aircrew some choice in their selection of route in the azimuth plane, such systems are called "Terrain Avoidance". This requires an azimuth instead of a vertical scan with the complete terrain profile being determined for each transmitted pulse. This is fortunately possible since the older Terrain Following systems waste a great deal of data by only extracting the boresight intersection point for each pulse when by suitable design of antenna, receiver and signal processing it is also possible to determine ranges and angles to the terrain over the complete elevation beamwidth.

One method of displaying the data from a Terrain Avoidance System is to paint on the Head-Up Display a series of transfer profiles for a sequence of range brackets say in 1500 m increments out to the longest range of interest, usually about 7000-10,000 m. It is still possible, in addition, to retain a terrain follow computation along track with such systems to provide a safety against ground collision no matter which azimuth route is selected.

2.5.2. Close Air Support

To provide support for troops by low level delivery of ordanance against ground designated targets lying close to their own position requires the highest standards in weapon delivery systems. Problems of target identification and acquisition loom large when the period in which visual sight of the target can be expected might only run to about 5 seconds.

The contribution which radar could bring to this situation has until recently been fairly small, essentially being limited to providing a range along boresight to the ground with the H.U.D. aiming mark and boresight slaved together. The accuracy to be expected from the radar with the low grazing angles involved, although not unduly high, has been compatible with the ballistic uncertainties of the weapons involved.

Current developments are taking a new direction. The radar can provide accurate navigation updates close to pre-planned targets and the high resolution techniques described earlier may also permit the target to be positively identified at some distance from the weapon release point. These are important contributions to the preliminary phase of an attack but to assist in the final phase where the target, if at all possible should be unequivocally identified, the need is for some form of automatic target acquisition.

This has created an interest in millimetric techniques since at these short wavelengths some form of imaging at least for certain forms of target such as armour is a possibility. When this ability is carried over into the weapon itself, a totally new capability emerges.

In this application the radar may act in a passive role relying on detecting small temperature differences between the target and its background or the aircraft radar may provide the illumination for a sensor carried in the weapon. The freedom from the effects of smoke which millimetric waves offer make the technique attractive for close support and the relatively small aperture required at short wavelengths allows sensors to be packaged within the confines of a weapon.

3. Air to Air

3.1. Airborne Early Warning (A.E.W.)

The task of an airborne early warning radar is very demanding: as a monostatic radar on a fixed-wing aircraft the prime function is to detect low-flying airborne threats and to report accurate target plots to a data processor for tracking, correlation with other data, threat assessment and so forth. This must be accomplished against the background of sea or land clutter, in the presence of various other forms of clutter and perhaps in the face of ECM. Very long range detection, as well as detection at shorter ranges, is required on aircraft and missile targets in all weathers: this means high mean power radars with large antennas in the frequency range 400 MHz to 6 GHz[3].

The provision of a large antenna has produced several quite different configurations in service, specifically under the fuselage for which size is limited mainly by ground clearance considerations, above the fuselage in a separate dome which allows the maximum size of conventional antenna, (here aerodynamic interference with the vertical stabilizer fin of the aircraft must be considered), and the FASS (Fore and Aft Scanner System) configuration which uses two antennas each surveying 180° in azimuth with a radome cross-section limited to a maximum of about the aircraft fuselage section. Several other configurations have been and are being studied; of particular interest is the use of 4 array antennas mounted in the wings and along the fuselage of candidate aircraft[4]. Such arrays can use distributed solid state transmitters and overall savings in weight and drag are possible. Consideration is being given to the use of a somewhat different pattern on reception compared to that used for transmission so as to place receive nulls on the worst transmit sidelobes: this should give a good two-way pattern and hence suppress sidelobe clutter.

The elevation pattern of the antenna also merits much attention. First because the short range coverage of low level (and high level) targets is basically determined by the elevation beamwidth, and secondly because high clutter levels can and do arise from short range ground clutter entering the receiver via the elevation sidelobes. There is a fundamental compromise between the short range coverage "hole" and clutter suppression. Consideration of the elevation pattern leads naturally to the question of target height determination: it is expected that a modern AEW will provide a smoothed height estimate on all tracks which will be derived from individual plot heights. The height is now basically derived from elevation angle of arrival and target range. The angle may be determined in several ways, including the use of a stack of beams in the vertical plane, or pseudo-monopulse multiple-beam techniques, or by providing an elevation scan to the radar beam. The latter form would now be undertaken electronically (such as E-3A) but in early AEWs was done mechanically as a separate nodding height-finder: the use of an auxiliary radar for height finding is now regarded as undesirable. The quite different "bounce" height-finding method used on E-2 relies on a reasonable level of reflection from the ground and moreover is ineffective at long range/low level; its use in any future system is quite uncertain.

Although in the past, pulse and ICW transmissions have been used for AEW, the current vogue is for pulse-Doppler waveforms either with or without pulse-compression. At 1 GHz and below a range-unambiguous PRF may be selected, but at higher RFs the clutter spectral width points to a range-ambiguous PRF used in conjunction with a ranging modulation. In all cases, there will of course be a target range-rate above which there are velocity ambiguities to resolve.

The duty ratio of the pulse-Doppler waveform is typically 1% and the peak power of the RF pulse about 1MW. At microwave frequencies, the coherent amplification may be provided by Klystrons or TWTs; CFAs may also be considered, but no firm plans to use such a tube in an AEW radar are known (the phased array antenna concepts would probably employ solid-state distributed transmitters). Power supplies for tube transmitters are best provided by a mixture of 400 Hz transformer-rectifier packs and inverters; technology for the latter is evolving and currently there is much interest in VMOS transistors. Grid modulation is preferable for the tube with a floating deck grid-modulator and optical coupling to the deck.

Microwave receivers may now be based on transistors, especially Field Effect Transistors, (FETs), such that adequate noise figures are achieved over wide bandwidths. Lower noise figures may be provided by the inclusion of parametric amplifiers but it is questionable whether the sensitivity improvement is worthwhile for systems specified to operate in an ECM environment. The protection of the receiver is provided by a combination of a gas T-R cell and low-loss solid-state stages such as PIN diodes. The capabilities of PIN protector technology has now progressed to the state where an all-solid-state protector design may be adopted.

The signal processing in an AEW radar is undertaken by a mixture of analogue and digital techniques. The bulk of the processing is in the latter category but matched filtering to each individual received pulse is more efficiently provided by an analogue processor such as a Surface Acoustic Wave (SAW) module. The use of block processing and multi-PRF ranging is recommended[5] for the detection of aircraft targets in heavy clutter. An efficient spectral estimation algorithm should be employed and this may be implemented by dedicated hardware or by firmware/software in a programmable processor. Radar detections following processing the block of data are made automatically by circuits which compare the signal in each range-velocity cell to that of neighbouring cells: unusually large signals are declared as possible targets and passed onto further processing. The detector is arranged to control the alarm rate on system noise. The detections arising from each block are then processed over a beam dwell time to produce radar plots; algorithms for estimating azimuth, height, velocity and so on are applied.

The target plots are passed to a radar tracking computer as the major contribution to building the air picture as perceived by the AEW System.

Several types of AEW System are in use today and in the years ahead these will undoubtedly be enhanced and updated to improve their capabilities and effectiveness.

3.2. Interception

The threat which radar equipped fighter aircraft attempt to counter has evolved from the high altitude intruder of the 1950's and 60's to the much more sophisticated terrain following multi-role aircraft of the present

day. Correspondingly the demands in terms of radar sophistication have grown so that competitive systems must offer a high capability in detecting airborne targets very close to the ground while operating at co-altitude or at higher altitudes. The problem which must be faced is that at co-altitude or in look-down geometries the earth's surface provides reflections which arrive at the receiver antenna at the same time instant as the target echo but with energy some 60dB larger. The radar system thus requires some 60dB or more "sub-clutter visibility". This can be obtained by pulse-doppler radars through frequency filtering if the doppler shift of the target is distinct from that of the ground clutter.

In the interception role the target is usually encountered in its frontal aspect so that the target doppler is appreciably larger than the ground doppler which originates from the motion of the interceptor alone. This implies that if the doppler band out to the maximum relative target to interceptor velocity is processed without aliasing, the target can be seen clear of all clutter and its detection is limited only by thermal noise. The requirement of no aliasing implies that the radar pulse repetition frequency (p.r.f.) be equal to the maximum anticipated doppler if complex sampling is used and be twice the anticipated doppler if the sampling is real. Systems which satisfy this (p.r.f.) constraint are known as "High Pulse Repetition Frequency" systems. The advantage of obtaining detection range on approaching targets which is determined purely by the power-aperture product is offset by the fact that targets which are receding produce dopplers which must compete with clutter from the antenna mainbeam or sidelobe ground echoes and the ranges obtained are much reduced. Additionally the requirement to avoid aliasing means that at microwave frequencies (typically X-band) the p.r.f. is so high that the unambiguous range is only a fraction of a nautical mile and some auxilliary modulation must be added to the transmitted pulse train for ranging purposes. Multiple p.r.f. ranging is seldom applicable.

Detections are formed into target tracks by a Track-While-Scan (T.W.S.) processor implementing an approximation to least mean squares estimation usually by means of a Kalman Filter algorithm. The target track information may be used to assess the target's threat potential or to evaluate the opportunity which may exist to attack it. Maintaining a regular scan pattern and T.W.S. processing gives minimum alert to the target that he is under scrutiny and may enable a closer approach to be made before evasive manoeuvres are encountered.

3.3. Combat

The principles applied to the interception problem fall short of solving the combat requirement where the fighter-target geometry may be quite arbitrary and in particular tail chase and turning flight have a dominant role. The clutter and target doppler bands are no longer disjoint so that some measure is prompted to make the target/clutter competition as favourable as possible. The most recent and most successful approach to this problem is termed "Medium Pulse Repetition Frequency" operation. The p.r.f. is reduced to a value about one tenth of the doppler bandwidth which produces aliasing deliberately. The sidelobe clutter is evenly distributed over the aliased frequency interval so that a uniform degradation is experienced. Main-beam clutter still occupies a fraction of the processed interval and denies detection over that fraction. The p.r.f. values used allow multiple p.r.f. ranging by transmitting a schedule of different p.r.f. values within the beam dwell and correlating the detections obtained across the schedule. Despite the best efforts in designing such a schedule some blind spots or blind zones always remain where target detection is impeded by main-beam clutter or other discrete components of side-lobe clutter. An example of a blind-zone diagram for this type of system is shown in Figure 3 where the dark regions indicate that the residual clutter density prevents detection.

Frequently Medium P.R.F. is found as an auxiliary mode in a radar which is primarily configured for a long range interception role using High P.R.F. However very careful design of the M.P.R.F. mode and the realisation of extremely low antenna sidelobes can lead to adequate interception performance using the M.P.R.F. waveform alone. This is an attractive development from the viewpoint of equipment mass, complexity and cost. A further advantage is that systems designed primarily for M.P.R.F. operation can use a low duty ratio T.W.T. which has the right properties for good air to surface operation. The best approach to providing full multimode capability at low mass lies in low duty ratio M.P.R.F. for interception/combat and Low P.R.F. for air to surface either under-running the tube or using simple digital pulse compression to maintain the mean power.

3.4. Weapon Considerations

The ability to attack a target at medium to long range by missile is one of the foremost arguments in favour of the airborne radar sensor. Current systems require that the radar enters a lock-follow mode dedicatedly tracking the victim target. The semi-active doppler missile either responds to the radar waveform or to a special illuminating waveform which is transmitted via the radar antenna. In either event the radar is blind to all activity apart from that along the single target sightline.

Future systems (AMRAAM) will use active missiles with launch information and guidance update supplied by the radar. The radar will be able to continue scanning while using T.W.S. processing to maintain target co-ordinates. Telemetry between the radar and missile may use some coding of the radar waveform or it may be more convenient to use an auxiliary missile telemetry sub-system. A major advantage of future systems will be the absence of a target alert due to the lock-follow mode or to special missile modulations when the missile is launched.

The necessity to direct guns in close combat poses a demanding requirement on the radar range and angle tracking loops. At close range the finite angular extent of the target induces glint or angular noise due to movement of the effective electromagnetic centre of the target with small changes in aspect. The combination of frequency agility which forces rapid glinting and digital filtering which can smooth the measurement sequence optimally afford a better capability to current radar designs than was achieved by earlier technology. Radome aberration remains as a significant source of error in this role and it will be practical in future systems to remove this by measuring and storing the radome aberration errors as a function of the antenna gimbal angles.

4. Air-to-Sea

4.1. General

Targets encountered in a maritime environment are either surface vessels or submarines. The aggregate of sensors carried by a military maritime aircraft (radar, sonar, ESM) can be optimised accordingly for an ASV (anti-surface vessel) or ASW (anti-submarine warfare) role.

The avionics suite of a modern military maritime aircraft consists of a group of sensors disposed peripherally relative to a central tactical system comprising a computer and universal TV-type displays on which tactical data and raw or processed video outputs from the sensors may be displayed individually or in combination.

The integration of the radar with other sensors in a total system, backed by extensive computational and display facilities, extends the radar capabilities by efficiently directing the gathering of radar data and by extracting the maximum of useful information from the raw data, by means of correlation with data from other sensors and by virtue of the greatly enhanced computing power now available. Significant advances in what might be described as the purely radar technology have also contributed to these new capabilites.

4.2. ASV

4.2.1. Detection

The main problem in maritime surveillance concerns the detection of small targets in heavy sea clutter. The use of frequency agile transmissions to decorrelate clutter on a pulse to pulse basis has become standard practice. The detection process must, however, take account of the variability of the clutter level and set the detection threshold at a level which achieves a specified Constant False Alarm Rate (CFAR).

CFAR processors may be either parametric or non-parametric, the distinction being that parametric systems have as their basis some assumption on the shape of the probability distribution of the clutter echo while non-parametric or distribution-free systems are based on rank-order hypothesis tests. Designs have till now been usually parametric and have revealed shortcomings in situations where the parametric assumptions are discovered to be unjustified. Non-parametric systems are based on rank-order hypothesis tests and have hitherto been little exploited because of the very high workload in forming the ranks.

One simplified approach (ref 1) involves the use of a Peak-Detecting CFAR system. This is more robust in the case of wrong assumptions in respect of clutter distribution than a conventional parametric system and this reduced dependence on the assumed clutter probability distribution makes it applicable to a wider variety of clutter conditions.

Alternatively, the advent of Contents Addressable Memory (CAM) in VLSI chip form will make possible the implementation of a fully overlapped rank-order system in two spatial co-ordinates, range and azimuth, which will significantly improve false alarm regulation and detectability for small targets in heavy sea clutter.

Clutter constitutes a genuine radar target and detection of wanted signals in clutter is achieved by comparing the probability distribution for signal plus clutter with that for clutter alone, setting an acceptable false alarm rate and required detection probability, which in turn defines the detection range. Thus the only way in which the signal to clutter ratio can be improved is by improving the ratio of wanted target echoing area to clutter echoing area, by means of smaller azimuth beamwidth and shorter pulse length. These, however, are assumed to be optimised initially for the required performance within the given installation constraints.

Detection performance in a noise-limited situation e.g. long range surveillance at grazing incidence, has been improved by the incorporation of low-noise FET amplifier r.f. front ends. Additionally, the use of flat aperture arrays leads to a greater aerial gain for the same swept volume. Microprocessor-based scan control permits optimisation of the scan limits and the scan rate to the search regime in use.

4.2.2. Location

Having detected a target, the military maritime aircraft may launch an on-board air-to-surface missile or it may fix the position of a designated ship target and relay that postion in geographical or ship grid co-ordinates back to a friendly missile fire control ship. The latter case may confer an over-the-horizon (OTH) targeting capability for launch of surface-to-surface missiles. To provide the fire control ship with target co-ordinates, the aircraft has the choice of computing target location in its own co-ordinate system and relying on being itself tracked by the fire control ship, or it can measure both target location and ship position and compute the target location in the ship's co-ordinate system. The latter system permits the fire control ship to maintain radar silence.

The enhanced computing power of modern avionics systems enables a large number of targets to be located and tracked while continuous scan is maintained. The location of targets can be computed in moving co-ordinate systems (e.g. relative to aircraft or fire control ship) or in a stationary grid system by stabilising positional co-ordinates against motion of the aircraft.

4.2.3. Classification

With OTH engagement there is a need to identify or classify targets, either to confirm previous intelligence that they might be hostile or to pick out the more important vessels. Investigations are being made into the possibility of ship classification using doppler signature or range versus amplitude profile, as described in section 5.

Doppler classification is considered to be an unreliable method since speed alone is not a good enough criterion. Antenna rotation is applicable to most ships of any size and under conditions of radar silence only the standard marine radar might be operative, so that all doppler signatures would be very similar.

Measurement of range profile requires good range resolution. Ideally targets should be near head-on in order to maximise target length. The basic requirment is for a radar to take snapshots of the vessels being scrutinised, otherwise maintaining radar silence, with a range resolution of the order of 50nS or less at a range of 50 nm plus on a frigate or larger target. The production of high resolution pulses is best achieved with some form of pulse compression. This fits in well with the modern trend away from non-coherent magnetron-based transmitters towards coherent Travelling Wave Amplifer-based designs. Additonally, enhanced resolution without loss of duty ratio at a modest pulse compression ratio can be obtained through use of the STRETCH technique.

The recent advances in computer processing and data storage capabilities permit the use of a large data base which facilitates comparison of measured range profiles with those of known threats.

4.3. ASW

4.3.1. Detection

In the ASW role, the main function of the radar concerns the elimination of surface contacts from the total of sonar contacts, so that those remaining can be presumed to be submarines. The underlaying of the radar display on the sonar display enables the observer to categorise contacts and to store this information in the

central computer.

An associated radar function is the detection of submarines which are surfaced or at periscope or snorkel depth. The small radar echoing area, particularly in conditions of high sea clutter, make periscope detection ranges very low, so this function is normally regarded as secondary.

The detection problems in the ASW role are similar to those in the ASV role. In the ASW role, detection is normally clutter-limited rather than noise-limited.

4.4. Missile Operation

Hitherto, it has been the practice to operate Air-to-Surface Semi-Active Guided Weapons with a fixed frequency or C.W. illuminator, with the missile head being locked up in frequency and, where appropriate, in range, before launch.

This limitation of a fixed frequency renders the system vulnerable to countermeasures. Frequency agility offers a solution against jammers other than fast response repeaters. It is, however, necessary to keep the missile receiver in step with the transmission. The move away from magnetron transmitters with mechanical and/or electronic tuning to power amplifiers driven from low power sources has made illuminator/missile alignment much simpler, since frequency selection can be made at low power and a random frequency code can be stored in both illuminator and missile and time-locked by means of the p.r.f.

5. Target Identification

5.1. Present Capability and Need

Modern aircraft are so expensive and so limited in weapon capacity that it is essential they attack the most cost-effective target with a high kill probability. Likewise, for lack of identification, they must not shoot up, or be shot down by, own forces in the heat of battle. The theoretical solution for all cases is to achieve a high degree of target recognition/identification, where the word target includes man-made or cultural features used for en-route navigation or an off-set attack. Target identification whether it be carried out by air-crew, or some form of automatic picture correlation depends on the recognition that its shape (or signature) and/or position in relation to other shapes matches prior knowledge. Tonal values appear to have a secondary effect except for ships. There is very little published data on radar recognition since, until the advent of in-flight video recording, it was difficult to carry out, or repeat, subject tests in a controlled statistical manner. Detection of aircraft, ships or vehicles in a clutter free , or clutter suppressed situation is now only limited by cost or weight but it may be crucial to an operation, in the more likely case of multiple targets, that the most significant one be identified from those detected.

Modern radars have the capability of accurately measuring the range and bearing of a target from an aircraft, its relative velocity, doppler signature and signal strength. Usually there must be some trade off in performance achieved in each parameter conditioned by cost and aircraft design restraints. One further method of man-made target identification is IFF (Identification of Friend from Foe). Since the earliest days of radar, IFF has been based on secondary radar system, which over the years has evolved from a single pulse system, through that of the multi-pulses, multi-code Mk X, Mk XII system to the proposed NATO NIS scheme. It is a sad fact that experience has shown the system is seldom 100% effective and that battle damage, crew fatigue, finger trouble and undetected failures all too often lead to a situation when an aircraft in confused close combat or late in returning to an entry gate or base can be shot down for lack of positive identification.

5.2. Air Targets

Leaving aside the special case of AEW, modern aircraft with a primary AI mode rely upon some form of pulse doppler radar to achieve detection and build up of tracks. Unless there is a vast superiority in speed (unlikely) there is no point in seeing targets which are not on a track that would lead to engagement in the next minutes. Thus future spatial separation may be the only requirement to identify the most cost effective target. In a cold war situation there is a severe range/sortie penalty when defence aircraft have to close to a visident position to confirm an intruder and it would be helpful if such could be identified at a longer range. Doppler signature of aircraft is no longer a reliable guide to resolve propeller driven or open turbine aircraft but a measurement of radar echo area (REA) coupled with speed and possibly length would discriminate many military aircraft from civil. Passive ESM is a last resort in the recognition of a threat but relies upon the target's co-operation and own knowledge, reception and recognition of his transmission, and might be too late even for self-survival. However, one might recognise a buddy operating in one's own r.f. band in time to break off an attack after the first engagements in a close combat melee. The recent advances in Surface Acoustic Wave (SAW) and Charge Coupled Devices (CCD) now permit fast frequency analysis and should not be neglected particularly if means can be found to improve the angular resolution of the sensor.

5.3. Ship Targets

Anyone can recognise on a radar display a large ship target on a calm sea though there can be a problem with small radars against small ships in even modest seas since the PPI noise like areas from targets can be lost in the sea patterns. Pulse compression radars have improved the performance against sea clutter and have permitted the use of monopulse resolution enhancement/improvement (MRE/MRI) adaptation of monopulse receivers to shorten the arc length thus subjectively improving the range/azimuth presentation of the ship. More sophisticated modes such as Doppler Beam Sharpening (DBS) or monopinch, which are now feasible through the use of coherent receivers and Fast Fourier Transfer (FFT) analysis can improve the azimuth resolution by a factor of 10 or more. Convoys and battle fleet forces present readily recognisable radar patterns within which it may be possible to sight the significant vessels purely from geometry, e.g. Carriers or SAW cruisers surrounded by area protection frigates; destroyers on the flanks and rear of convoys. Smaller naval forces such as Krivak and Nanuchka lurking amongst fishing vessels are unlikely to be identified by pattern recognition.

Experiments have been carried out on the classification of military vessels by doppler signatures of scanning antennas or high resolution range amplitude profiles. Whilst it is feasible to obtain useful doppler data from ground based radars against naval vessels in relatively smooth water if the large surveillance antenna is turning, the resolution problem is formidable when the sensor is in an aircraft and the ship is rolling and pitching in a rough sea. The new FFT techniques with longer integration times and narrow filter bandwidth analysis offer some improvement in performance. Detection of the standard marine radar antenna is of little value as an identification feature as it is carried by most sea-going vessels.

Range/amplitude profile does however seem to offer some potential given a range resolution of 10-20 feet. There are some 17,000 registered merchant ships of length greater than 200 feet compared with some 600 naval ships of frigate size or larger. Fig 4 illustrates a histogram of deck length in increments of 10 feet for merchant ships and the navies of the UK, USA and USSR. Most of the smaller ships are general cargo and most of the largest ships are rear engined bulk or fuel carriers all of which are likely to have an in-line radar echoing area (REA) profile significantly different from that of frigates or cruiser/carriers. With the larger rear-engine ships the massive bridge/accommodation structure either forward or aft will create a long radar shadow when the ship is viewed +15° off line and with the comparatively clean deck these big vessels may be shortened by some 100 feet radar wise. Some USSR warships have a very clean deck aft of the main mast and might appear short when viewed from the head. When viewed from the rear the quarter deck should indicate the stern but the high bridge and tiered radar stations may screen the bow.

Frigate lengths lie between 270-300 feet USSR, around 360 feet UK and 415-440 feet USA. None of these correspond to the peak distributions in the merchant ship histogram, and on a probability basis might be expected to be resolved by range profile, since even though the USSR frigate lengths lie close to the shoulder of the peak most of the vessels concerned are rear engined with bridge aft. Destroyers cover a length range 390-470 feet. These values overlap the largest peak in the merchant ship histogram with most of the vessels of a composite profile with bridge forward and accommodation, funnels and superstructure occupying some two thirds deck length, and would be more difficult to discriminate from destroyers (except when at high speed). Cruisers lie within 360-720 feet with the bulk of the Soviet ships in two groups around 520 feet and 650 feet. The shorter vessels could be confused with composite merchant ships but the longer ships compete, mainly with rear engine bulk/fuel carriers. Moreover most USSR cruisers have massive radars and sensors in tiers on the central masts/funnels and should provide a unique range amplitude profile compared with other NATO and merchant ships.

Apart from Invincible class (677 feet) and Kiev class (925 feet) most aircraft carriers are in excess of 1000 feet. Though most of the USA amphibious carriers are under 900 feet these are unlikely to be present except in a land assault operation. The majority of merchant ships in the carrier range bracket are rear engined with little superstructure and hull low in the water. Even the derricks, services or hatches along deck are unlikely to produce an amplitude profile compatible with an aircraft carrier. Thus it seems probable that given adequate range and amplitude resolution and a favourable track it should be possible to correlate observations with stored profiles of ship lengths to at least discriminate a naval vessel from merchant shipping and maybe designate probable class.

5.4. Vehicle Targets

Detection of military vehicles demands the use of a high resolution radar in order to reduce the effects of ground clutter and preferably the addition of Moving Target Indication (MTI) to concentrate attention on moving targets. Discrimination from civil vehicles and identification as to type presents a greater problem though an armoured thrust might be recognised purely from the geometric in line or box formation. Special field sensors can detect a difference in some types of tracked vehicles by the doppler signature of the platelets. Airborne spotlight radars with DBS are now being developed to achieve such discrimination since clearly it would be advantageous if a strike or recce aircraft could pick out GW or tank vehicles from others beyond visible range even though, for weapon kill accuracy, the attack would have to be completed visually.

5.5. Land Targets or IPs

Aircraft survival requires operation at low altitude in enemy territory. As a consequence the radar view seldom matches the classical plan drawn from a map or photo recce due to range screening from woods, small hills as well as buildings. In the past navigation fix points (IPs) were chosen (and surveyed) to be both visually and radar identifiable (e.g. isolated towers, lighthouses, water boundaries, corners or woods etc), since given adequate navigational aids they could be quickly recognised without ambiguity. Flying at an altitude of some 300 feet the unmask range of typical interdiction targets (airfields, rail junctions, POL sites) lies around 9 n. miles in undulating terrain falling to 5 n. miles in hilly or mountainous areas as compared with the flat earth range of 20 n. miles. Post flight analysis of radar recorded flights suggests that though detection was possible at some 70% - 90% of unmask range a trained operator would not confirm (identify) the target until 50% or less of unmask range. Reasons for the delay are not clear but until the advent of DBS the azimuth resolution of airborne radars was much poorer than that in range and operators had to rely more upon cross track line features and patterns for navigation and target cues. Even the latest automatic picturematch for infra-red (IR) look down sensors, suggests line features within 45°-90° off track offer the highest accuracy.

Use of DBS with a potential angular accuracy of 1mR might increase the number of IPs that could be chosen for en-route navigation or off-set attack, but will not improve resolution along track. Very short pulse or high compression radars can effectively improve angular separation of a number of discrete targets (e.g. tanks, aircraft shelters) by the use of MRE since only one will appear in the azimuth/range cell, and thereby will improve the chance of identification by pattern association. Some military targets can be recognised without the need for high resolution. For example GW sites, airfields and POLs tend to have fairly standard well separated discrete high constrast REAs, which to an experienced radar operator provide cues. Even on an airfield the control tower, ILS installation and defence weapon sites should be sufficiently isolated in open ground to give a head-on (or off-set) for an attack on less obvious more significant targets.

5.6. IFF

The conventional secondary radar IFF system used in NATO and a similar method in Civil Aviation is intended to provide positive identification of aircraft, military ships and in some cases ground sites or vehicles. In practice the present IFF system has disadvantages since despite elaborate coding it may be open to spoofing, it permits the enemy to home on to replies as did their fighter to our bombers in the last war, and because of battle damage or unreliability it is not fail-safe against quick reaction missile systems. None the less NATO has decided to embark upon a new NIS system to a draft STANAG agreed in 1980. Major aims of the new system would be to achieve a dramatic increase in security, reduction to single target interrogation coverage and extension of single ground units in the battle field. One option is to use the primary airborne sensor (radar or laser) to interrogate on the same or related bands using a cryptic Question/Answer (QA) code only when the target is in main beam, the latter replying on one of several standard frequencies only when it is known it is in interrogation main beam. Such a system works beautifully on paper and on the bench but it is questionable whether, after the first battle, field maintenance and communication problems will be up to keeping the daily or

hourly code selections and electrical units reliable. Hence of the three limitations mentioned earlier only that of spoofing becomes less likely. The principle of within beam interrogation provides an unwelcome and serious challenge to the design of future airborne radars. The antenna will have to be compatible with sidelobe suppression and responsive to both radar and NIS frequencies. The transmitter should be capable of transmitting the long train of coded pulses interspersed with the radar pulses or a separate unit will have to be duplexed into the antenna systems with a risk of damaging the primary radar receiver. The time taken for the QA code might limit the rate of antenna scan and would presumably proscribe the use of fast diverse electronic scan. Even if NIS solves the close combat air-to-air identification problem it cannot help in the discrimination of naval/merchant shipping or ground IP/targets. Thus there will be a need to include in the design of new airborne radars the highest possible range, angle, doppler resolution and possibly digital picture/target correlation compatible with cost, volume and weight restrictions if the aircraft system is to be successful in a multiple target situation.

REFERENCES

(1) John C. Kirk Jr.

A Discussion of Digital Processing in Synthetic Aperture Radar.
IEEE Trans. Aerospace and Electronic Systems Vol. AES-11 No. 3 May 1975

(2) John C. Kirk Jr.

Motion Compensation for Synthetic Aperture Radar.
IEEE Trans. Aerospace and Electronic Systems Vol. AES-11 No. 3 May 1975.

(3) J. Clarke

Considerations Relating to the Design of Modern AEW Radars.
MEDE EXPO 78 Conference Proceedings, Wiesbaden, October 1978 pp 303-312.

(4) J.K. Smith

Space Density Tapered Arrays for Airborne Surveillance Radar.
Proc. IEEE Int. Radar Conference, Washington, April 1980 pp 263-270.

(5) J.M. Colin

Radar Design for Detection in Clutter.
Proc. IEEE International Radar Conference, Washington D.C. 28-30 April 1980 pp 317-321.

ACKNOWLEDGEMENTS

The author wishes to draw attention to the valuable contributions to this paper from the following people:-

Dr. J. Clarke, R.S.R.E., Mr. D. Farquhar, Mr. J. Roulston of Ferranti plc and Mr. E.K. Williams, consultant to Ferranti plc.

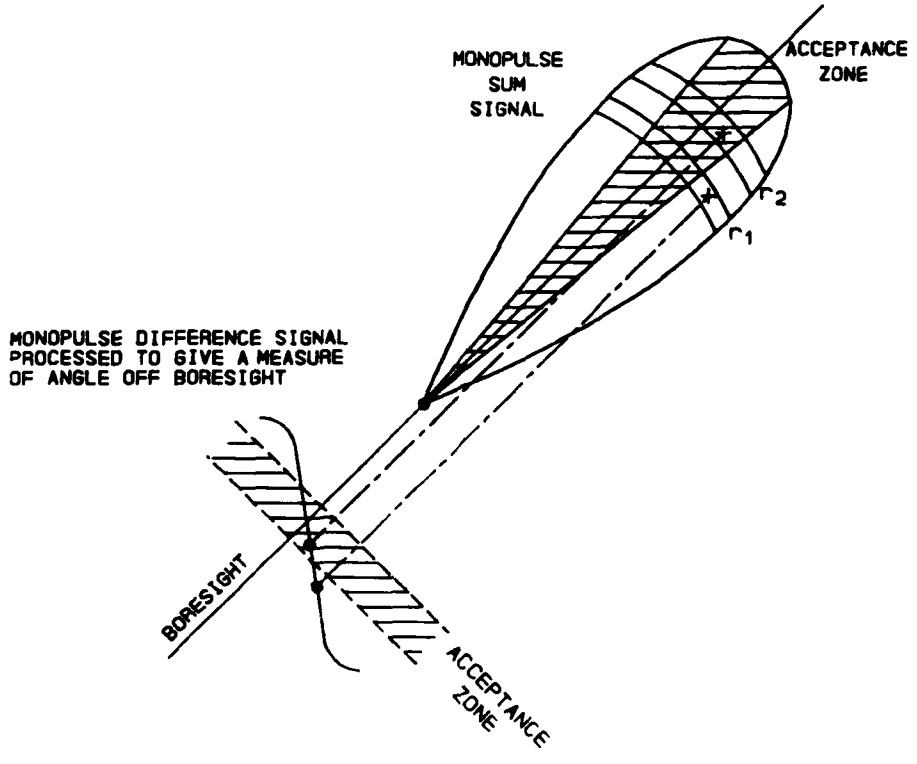


FIG1 MONOPULSE RESOLUTION ENHANCEMENT (MRE)

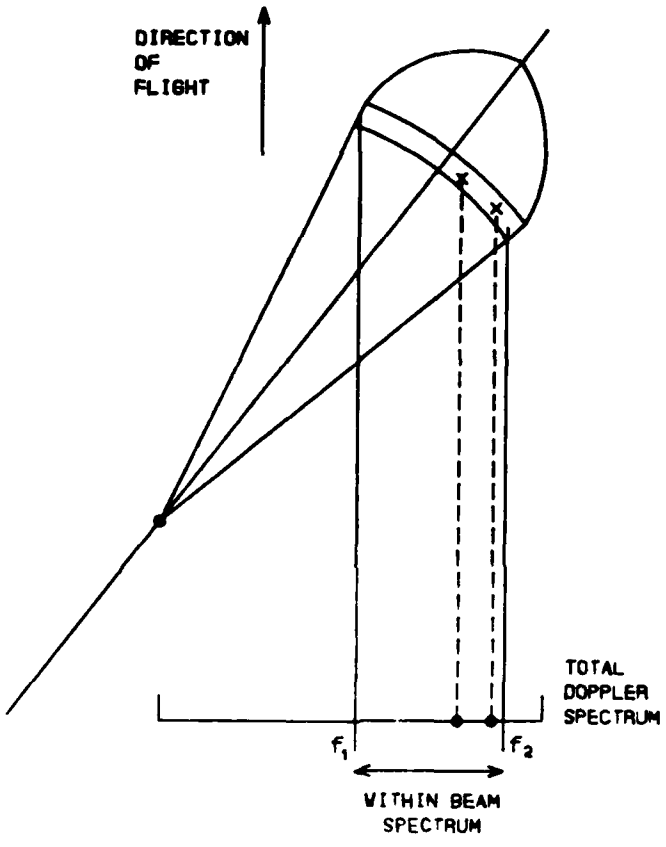


FIG2 DOPPLER BEAM SHARPENING (DBS)

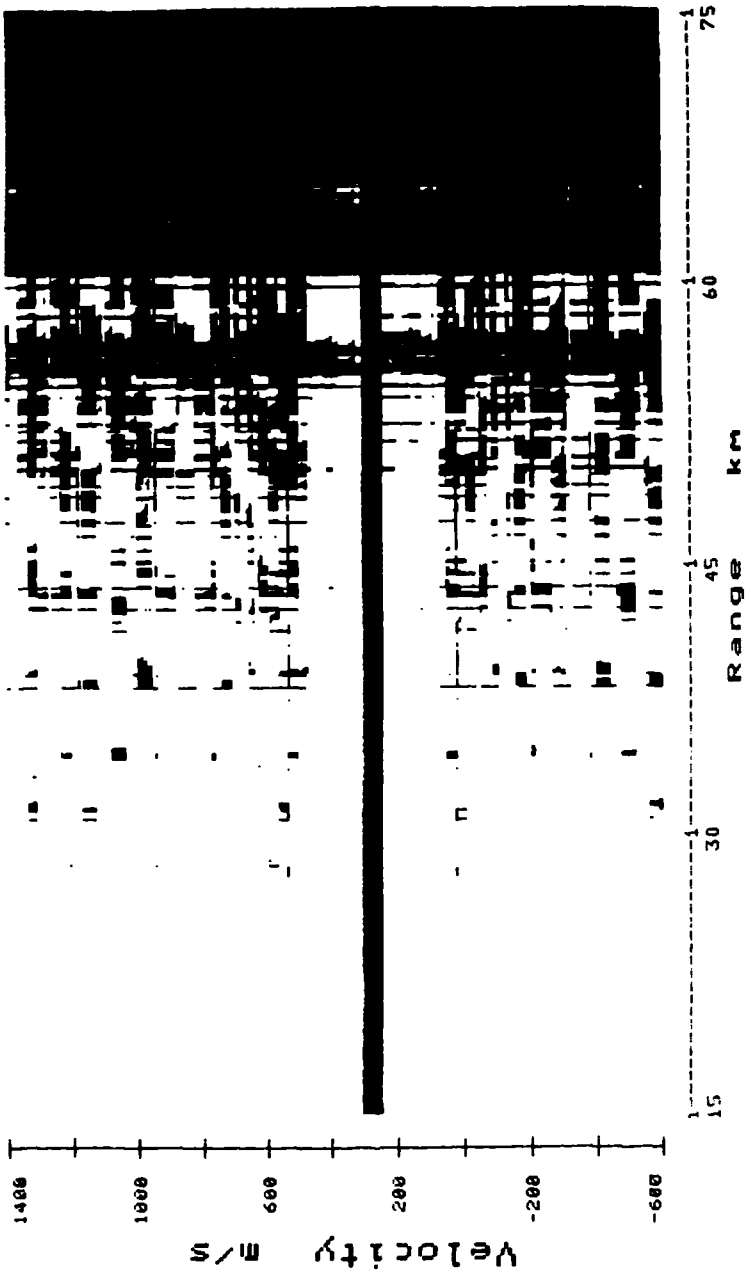


FIG 3 Blind Zone diagram for M.P.R.F. radar. The dark regions indicate single dwell detection probability below 50%.

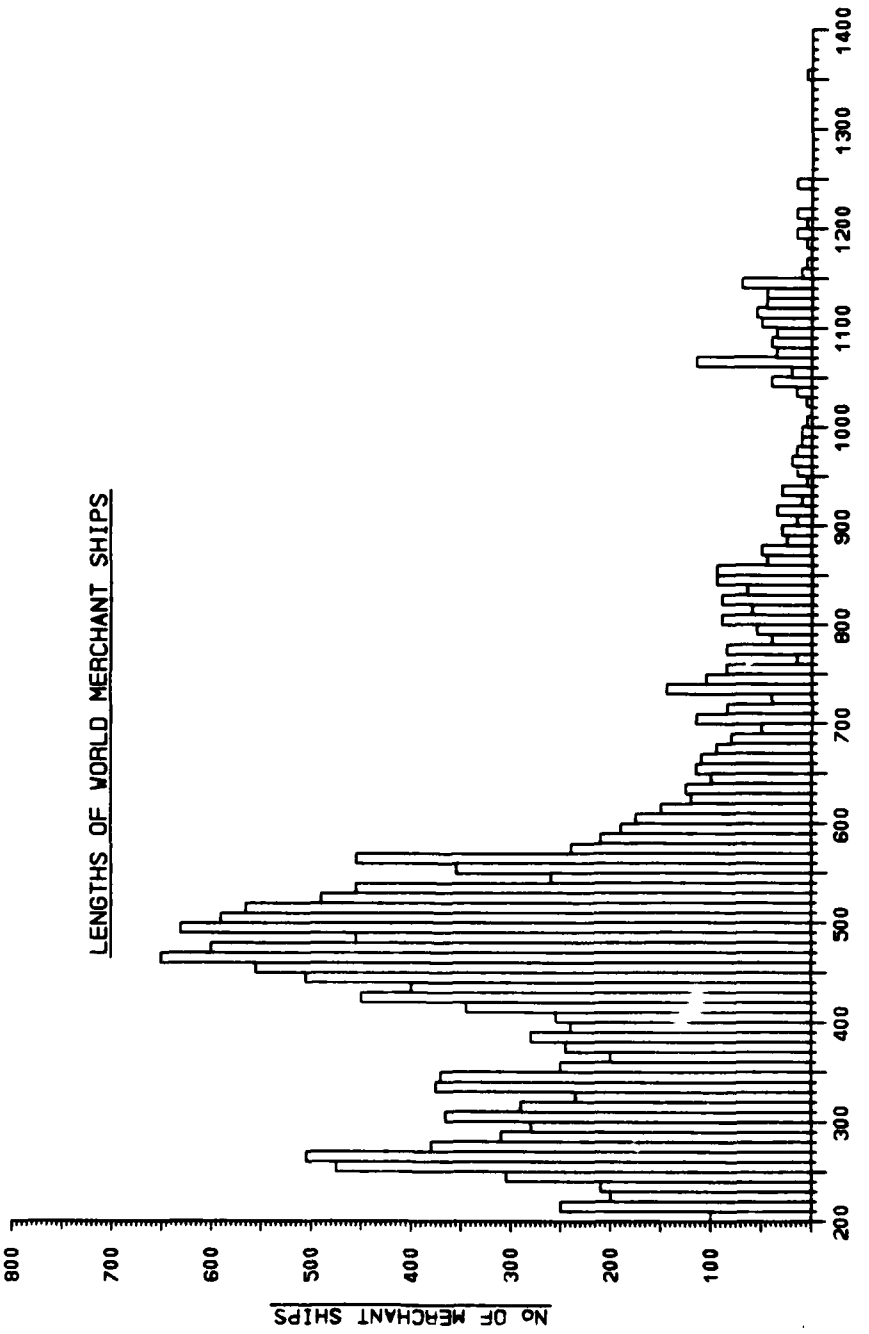

LENGTHS OF WORLD MERCHANT SHIPS
No OF MERCHANT SHIPS
800
700
600
500
400
300
200
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
SHIP LENGTHS (ft)

FIG 4

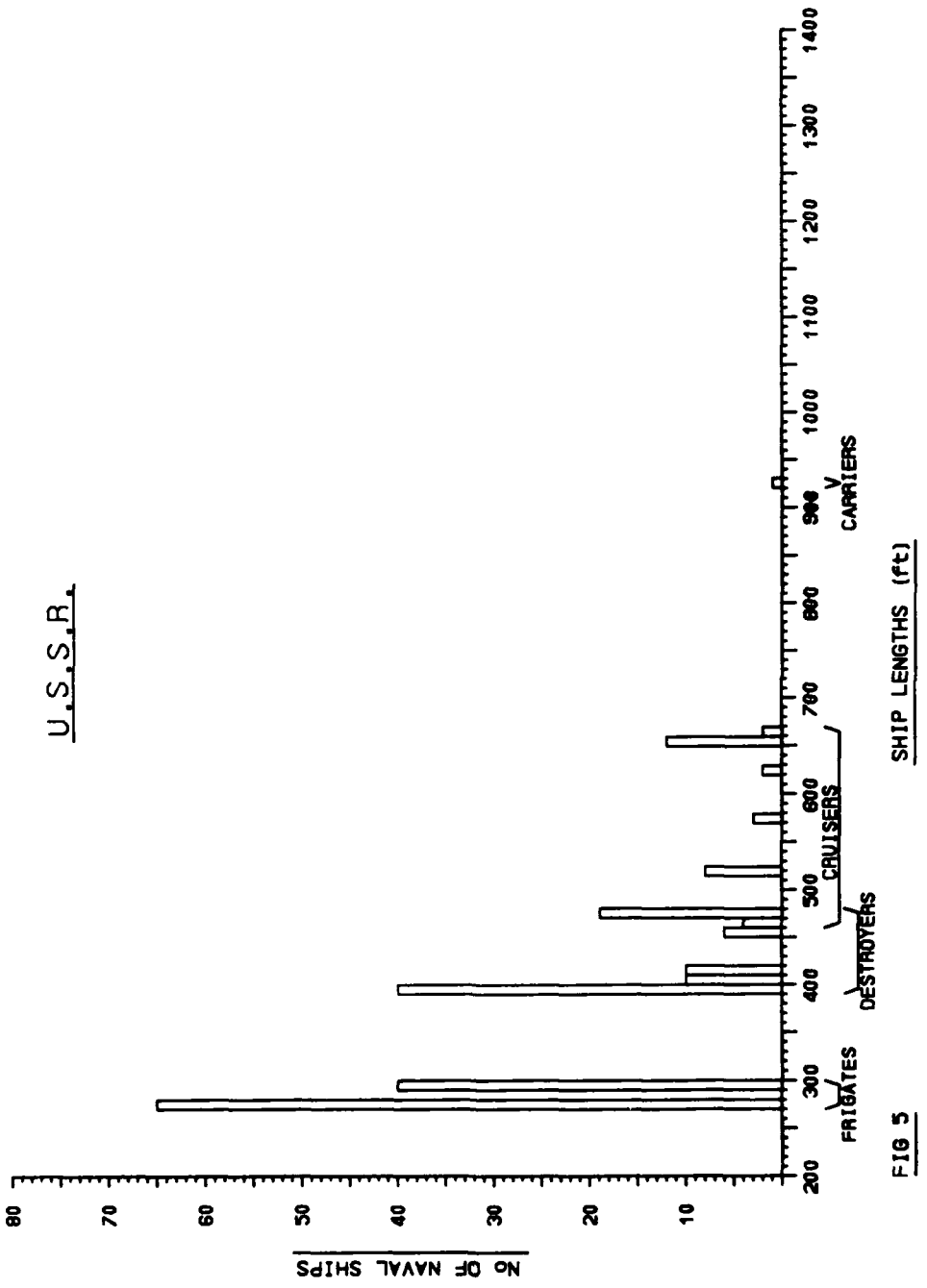

U.S.S.R.
No OF NAVAL SHIPS
80
70
60
50
40
30
20
10
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
SHIP LENGTHS (ft)
FRIGATES
DESTROYERS
CRUISERS
CARRIERS

FIG 5

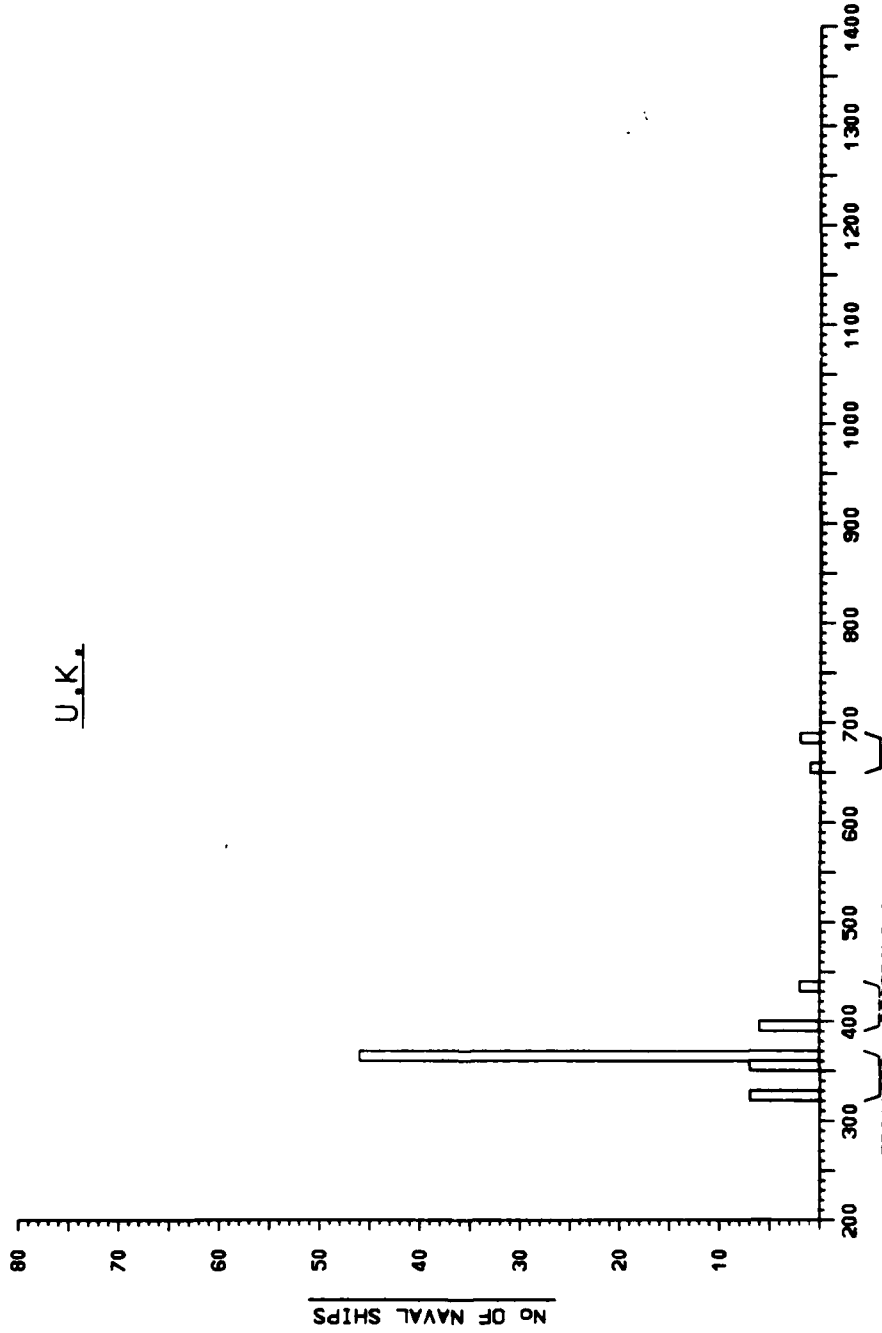

U.K.
No OF NAVAL SHIPS
80
70
60
50
40
30
20
10
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
FRIGATES
DESTROYERS
CARRIERS
SHIP LENGTHS (ft)

FIG 6

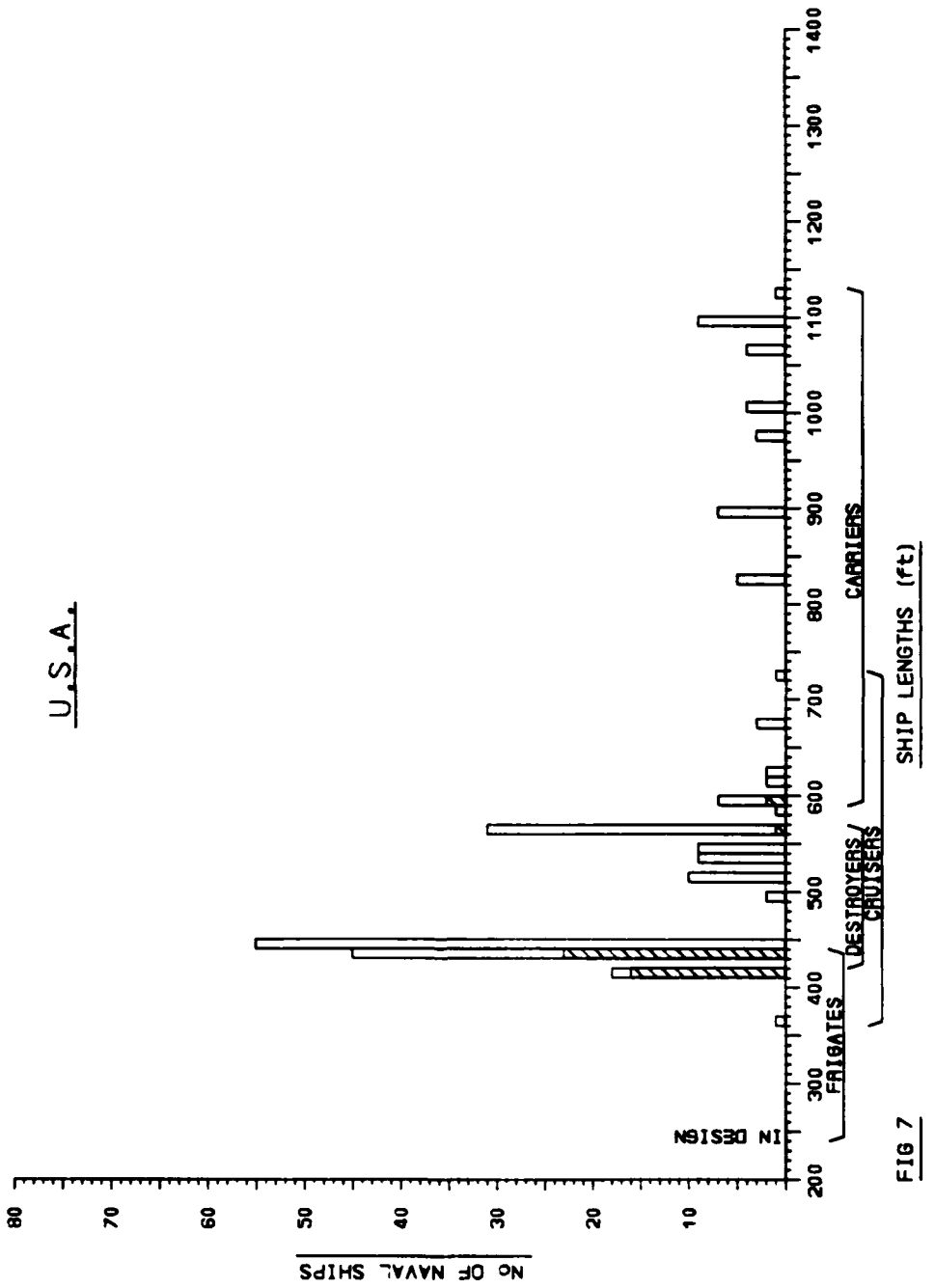

U.S.A.
No OF NAVAL SHIPS
80
70
60
50
40
30
20
10
IN DESIGN
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
SHIP LENGTHS (ft)
FRIGATES
DESTROYERS
CRUISERS
CARRIERS

FIG 7

INTRODUCTION TO OPTICAL RATE SENSORS

by
Uwe K.Krogmann
Bodenseewerk Gerätetechnik GmbH
7770 Überlingen, GE

SUMMARY

There is a wide range of guidance and control application for which rate sensors with different requirements are needed. The accuracy requirements cover 7 to 8 decimal powers.

So far these applications have mainly been using electromechanical gyros of which the dynamically tuned gyro (DTG) has recently gained considerable importance.

In order to fully utilize the cost potential of inertial systems in strapdown configuration and thus to usefully employ inertial technology in further fields of application the use of simple sensors involving lower costs without performance degradation is presently a very important challange.

The active Ring Laser Gyro (RLG) has proven its performance potential recently. It is however still an expensive sensor.

Due to their suitable measurement model in a dynamical environment (e. g. acceleration insensitive) and due to the favourable cost potential the passive optical gyros are expected to obtain the breakthrough in this field, in the near future.

In case of mechanical gyros the reference for rate measurements is the fact that the spin vector remains constant. In case of optical rate sensors the inertial reference is the velocity of light which remains constant independent from the motion of the case.

The basic principle of all optical rate sensors is the Sagnac Effect. It is based on the measurement of optical path differences between two counter-rotating beams in a measurement ring which rotates with respect to the inertial space.

In case of this type of measurement arrangements with dimensions acceptable for inertial sensors the simple Sagnac Effect leads to unfeasible measurement resolutions assuming low rates (e. g. 0.01°/h).

The employment of the Sagnac Effect for rate measurements requires an increase of the resolution capability. This is achieved by

- a multi-turn of the light-beam in low-loss light-conductive fibre coils (Sagnac Interferometer)
- the multiple transit of beams in only one optical ring (passive ring resonator)
- the integration of a laser medium into a mirror ring arrangement (active ring resonator).

Starting from the Sagnac Effect the physical functional principle of these three realizations will be presented here.

The first two realizations are passive sensors. They are characterized by the fact that an external light source radiates into an interferometer or ring resonator. Depending on the angular rate subjected to the arrangement the measurable light intensity of an interference fringe pattern changes. This change is proportional to the angular rate and is determined basically in an analogue-type measurement.

Unlike the above two mechanizations the third mentioned realization leads to an active sensor which integrates a laser medium into a mirror ring arrangement. Due to the thus formed ring resonator the Sagnac path difference which occurs in case of two counter rotating beams is converted into a high-resolution frequency difference which can be immediately determined in a digital type measurement.

In all mentioned designs the application of the difference measurement principle is decisive (2 counter-rotating beams, 2 resonances for counter-rotating beams).

1. RATE SENSOR APPLICATION IN GUIDANCE AND CONTROL SYSTEMS

1.1 The Notion "Gyro"

The problem of performing angular rate measurement automatically leads to the notion "Gyro". As shown in fig. 1-1 this notion has a different meaning in which field of technology it occurs.

The earth itself shows almost perfect gyro characteristics the slow polar axis drift being due to a precession motion which is caused by the rotation of the earth and disturbing torques.

Since it is here dealt with inertial technology the notion "Gyro" shall include any unit which measures angular rates or attitude changes per unit of time.

For defining direction, directional changes, or rotation the fixed stars are the spatial or inertial frame of reference (see fig. 1-2).

The observation of the stars is consequently a straight-forward means to measure inertial rotation. Indeed it is well known that this is the basis of celestial navigation.

As far as terestrial navigation is concerned stellar observations are very often not feasible or even impossible.

Moreover, special instruments are required which have to be capable of sensing rotation with respect to inertial space within the closed environment of a vehicle and without the need for external communication (autonomous navigation, guidance).

In 1851 Foucault designed the first instrument for sensing the earth's rotation without applying celestial references. It consisted of a very long (approx. 70m) wire pendulum carrying a heavy mass shere. It hung from the dome of the Pantheon in Paris. The oscillation direction of such a pendulum remains fixed while the earth below it turns thus enabling the measurement of earth's rotation.

The first spinning rotor gyroscope was again designed by Foucault in 1852. He applied the - at that time already known - effect that spinning bodies possess the feature of keeping their rotational axis fixed with respect to space if no disturbance torques are effective. With this first free gyro he again observed earth's rate. It is from Foucault's instrument that the notion "gyroscope" originats:

- gyros, which means rotation
- scope, which means to view.

Approximately seventy years ago Hermann Anschütz of Germany and Elmer Sperry of USA designed - independently from each other - the first practically applicable gyroscopic instrument, the gyrocompass.

Since that time many different gyroscopic instruments have been designed before the first applicable inertial navigation system became available immediately after world war II. It was designed by a team headed by Charles Stark Draper in the USA. The historical steps in the development of mechanical rate sensors are summarized in fig. 1-3.

In the meantime inertial systems and attitude and heading references systems have experienced trimendous technological progress. The electromechanical gyros mainly applied within these systems so far are based on the facts that

- the spin axis of a free, undisturbed and perfectly balanced rotating mass keeps its direction fixed to inertial space.

- The direction of the spin axis can be changed deliberately by application of a precessing torque.

A break-through of optical gyros (rate sensors) is expected in the near future which could have a considerable impact on the inertial sensor market.

1.2 Rate Sensor Requirements

Within the scope of guidance and control there is a wide range of application for rate sensors. As shown in fig. 1-4 the requirements of the different applications cover a range of 7 to 8 decimal powers. The highest requirements are to be met by the highly-accurate azimuth self-alignment by measuring the components of the earth's rotation and for the guidance of ballistic missiles. The latter application requires furthermore the extreme availability and reliability of the sensors.

The lowest accuracy requirements are to be met in connection with flight control and stabilization systems as well as short range missiles and guided projectiles. The rate sensors for flight control systems have to meet very high reliability requirements, too.

Between these two extremes there is a wide range of different applications and requirements. This wide range of applications and accuracies has lead to the effect that today many different rate sensor types are used which are based on different physical principles. Some of the presently employed gyro types are indicated in fig. 1-5 together with newly developed designs.

When considering the requirements of the different applications presented in fig. 1-4 and by comparing them to the projected rate sensor potential as described in fig. 1-6 the following can be stated:

- In the recent past up to today the dynamically tuned rotor gyro (DTG) has covered a large part of the application range although it has never reached the required highest accuracy and although its costs are too high considering the extreme inaccuracy range of applications.

- A further increase in the accuracy of the DTG cannot be expected for the future.

- In the future the ring laser gyro (RLG) and the passive ring interferometer can basically cover the whole range of applications including the highest accuracy values. Here especially the ring interferometer is expected to bring about a decisive break-through as far as cost versus performance are concerned.

From the present point of view the optical rate sensors are extremely important due to their basic potential for a wide range of applications within the scope of future guidance and control systems /53/.

When looking at the rate sensor potential it has to be noted that there are some phenomena which limit the rate measurement accuracies obtainable on the earth. In fig. 1-7 some of these phenomena are indicated. A compensation of the three first-mentioned effects is possible because they are deterministic in nature.

1.3 Optical Rate Sensor Penetration - why Optical Rate Sensors ?

Considering the evolution of inertial systems which represent a major component in autonomous navigation, flight guidance and control systems the future importance of optical rate sensors is further reinforced.

The inertial measurement units prevalently applied within inertial systems in use today are realized as gyro-stabilized gimbaled platforms. On the stabilized element (cluster) of which the inertial sensors (gyros, accelerometer) are located, as can be seen from fig. 1-8. Due to the use of electromechanical gyros, electromechanical accelerometers as well as of a complex electromechanical gimbal system these units involve high costs.

Due to the development of high-performance signal processing hardware inertial technology could realize the transition from platform type systems to strapdown systems. In the strapdown version the sensors are mounted body-fixed and are thus subject to the full rotational dynamics of the carrier vehicle. This rotational dynamics which may cover a range of 0.01deg/h up to some 100deg/h establishes considerable requirements to be met not only by the sensors but also by the signalprocessing hardware of strapdown systems. Here inertial technology could benefit from the rapid development in signalprocessing hardware (VLSI technology) in the commercial field. Due to the achieved miniturization levels it has become possible to realize strapdown systems with extremly high-performance computers of affordable size and cost.

As far as the sensors within such systems are concerned things are still a little bit different. At first the procedure had been to transfer the electromechanical sensor technologies that had proved to be good for platform applications to sensors for strapdown applications. But these sensors are still very complex electromechanical high precision components. However in the course of the last ten years these electromechanical gyros have been developed into rate sensors providing adequate precision and realiability values.

For the application in advanced strapdown systems in which the sensors are - as already mentioned - subject to the full translatory and also rotational dynamics the mechanical gyro - as a result of its complex measurement model - requires a considerable static and dynamic error compensation, which inturn requires calibration in regular intervalls.

In order to further reduce costs and thus increase the cost-effectiveness and affordability of advanced inertial strapdown systems efforts had to be started to eliminate these expensive mechanical sensors.

As to the gyros considerable progress has been made in the last years where inertial technology could largely benefit from the use of optical techniques developed in the

commercial communications field and it will benefit even more in the future (fiber optics and corresponding components, laser, integrated optics, etc.)

When the further miniturization within signalprocessing hardware (VHSIC - Very High Speed Integrated Circuit) which is being advanced particular in the USA will be taken into account it will soon be possible to realize all solid state handy inertial strapdown systems which can be used at reasonable costs in a wide range of applications meeting the most differing requirements. Due to their advantages as compared to electromechanical components the optical rate sensors will play an important role here. As an example the main advantages of the ring laser gyro which among optical rate sensors has achieved the most mature design status are compiled in fig. 1-9.

These laser gyro atributes not least are due to the different measurement models of the ring laser gyro and today's most commonly used electromechanical gyro, the dry tuned gyro (DTG). These two measurement models are compared to each other in fig. 1-10. The basic function of the ring laser gyro is described in detail in paragraph 3.3.

Due to the elimination of the rotating mass in the laser gyro the acceleration sensitive as well as dynamical terms do not occur in the measurement model of this gyro type. Thus the performance of a ring laser gyro can basically be described using only the following three parameters:

- scale factor error
- misalignment angle
- bias drift, drift noise.

Even when it is taken into account that drift noise, weight and volume of a ring laser gyro (RLG) are larger than those of a classical mechanical gyro of comparabel performance grade the potential advantages of the RLG listed in fig. 1-9 together with the superior measurement model shown in fig.1-10 are considered more important. This particularly applies for strapdown systems with their hostile dynamic environment. Just recently the ring laser gyro has proven its performance potential, however, it is still an expensive piece of equipment.

An all solid state strapdown inertial system utilizing ring laser gyro technology is shown in fig. 1-11. Remembering the ring laser gyro attributes mentioned such a system offers significant benefits to the user. Among these are the important features listed in fig. 1-12.

What has been stated so far about strapdown systems in general and optical rate sensors in particular will help to understand the market potential of inertial systems as it is presented in fig. 1-13.

The following essential facts result from this figure and are compiled in fig. 1-14.

- Due to the use of advanced sensor and signal-processing techniques and due to the thus resulting realization of low-cost systems there will be an increasing market for inertial systems covering the whole range of applications.

- The market share of systems in strapdown configuration will considerably increase in the future.

- The expected increase in the market share of strapdown systems results among other things also from the optimism which can presently be found in connection with the use of costeffective optical rate sensors incorporating the advantages mentioned before for such sensors.

- Thus the use of optical sensors is expected to increase parallel to the number of employed strapdown configuration systems.

However, fig. 1-13 shows also that the dynamically tuned gyros (DTG) will keep a certain share in the market. This is due to the facts that

- these gyros have achieved an advanced state of technical maturity and reliability.

- The achieved accuracies sufficiently meet the requirements of many applications.

- There is a certain "logistic inertia" which favours the further use of introduced and proven products offering well established logistics.

In fig. 1-15 the most important features of the DTG are summarized and compared with those of the ring laser gyro.

Due to the continuous advances made in the development of lasers, optical fibers and integrated optics technology there will be light sources, optical transmitters and signal transmission equipment available which will considerably influence the future design of optical rate sensors.

It is not to be overlooked that a decisive break-through in the development of passive optical rate sensors (lowest cost) could within a very short period of time change the market shares considerably in favour of these sensors, a fact which would all of a sudden extremly expand the field of application of inertial technology.

2. REVIEW OF OPTICAL RATE SENSOR FUNDAMENTALS

2.1 Basic Function of Optical Rate Sensors

Inertial sensor technology is finally based on the inertia which a mechanical mass or electromagnetic radiation oppose to any motion or movement due to the mass energy equivalence /55/.

Schultz - Du Bois explained in /55/ that according to this the radiation energy stored in a ring resonator is equivalent to a heavy mass ring within a round frictionless bearing surrounding case. With this electromagnetic (optical) analogy to mechanical instruments he provides an alternative interpretation of optical rate sensing and shows the reasons why it is basically possible to employ electromagnetic radiation in inertial technology (see also fig. 2-1).

As shown in fig. 2-2 and already mentioned in paragraph 1.1 the rate measurement reference of mechanical gyros is based on the fact that the spin of a rotating mass remains constant whereas in case of optical rate sensors the inertial reference is the velocity of light which remains constant independent from the motions of the case.

The basic principle on which all optical rate sensors are based is the Sagnac Effect. This effect involves the observation of nonreciprocal i. e. differing optical path lengths as a function of the direction and amount of rotation.

This phenomenon directly results from the postulate of the theory of relativity concerning the constancy of the velocity of light.

2.2 The Sagnac Effect

In order to be exact the Sagnac Effect - i. e. the evidence of a rotational motion using optical means but no external reference - would have to be dealt within the formalism of the general theory of relativity since any rotational motion represents an accelerated motion.

In principle a difference would have to be made between whether the signal detection is done by a stationary or by an also rotating observer. As shown in the literature /2/ the first approximation leads to identical results in both cases.

The easiest way to derive the Sagnac formula is to assume a circular light path which is rotated about an axis perpendicular to its area with the constant angular rate Ω .

Two counter-rotating beams are assumed to start from point C in the path of the arrangement shown in fig. 2-3. The ccw light beam (b) hits this point again at the point C' of a reference system which is assumed to be stationary; the distance between C and C' is ΔS_b . The cw light beam (a) hits its entering point in C'' which is ΔS_a apart from C.

For a stationary observer the propagation times for both beams are:

$$\tau_b = \frac{2\pi R - \Delta S_b}{c} = \frac{\Delta S_b}{R \cdot \Omega} \qquad (2\text{-}1a)$$

$$\tau_a = \frac{2\pi R + \Delta S_a}{c} = \frac{\Delta S_a}{R \cdot \Omega} \qquad (2\text{-}1b)$$

$c \triangleq$ velocity of light in a vacuum
$\Omega \triangleq$ angular rate of the arrangement

τ_a is obviously unequal τ_b . The elimination of ΔS_b and ΔS_a leads to:

$$\Delta\tau_N = \tau_a - \tau_b$$

$$\Delta\tau_N = \frac{4A\Omega}{c^2 - (R\Omega)^2} \qquad (2\text{-}2)$$

where $A = R^2 \cdot \pi$ which is the area surrounded by the beam.

Time dilution:

Since this time interval is observed in a stationary system it is delayed by a factor γ with respect to an also rotating observer (time dilution). I. e. the time interval $\Delta\tau_R$ measured on the also rotating beam splitter is

$$\Delta\tau_R = \Delta\tau_N/\gamma \tag{2-3}$$

or

$$\Delta\tau_R = \frac{4A\Omega}{\gamma(c^2-(R\Omega)^2)} \tag{2-4}$$

In case of the Galilei Transformation which is based on the postulate of absolute time γ would have to be 1. If on the other hand the dilution factor of the Lorentz Transformation is used

$$\gamma = \frac{1}{\sqrt{1-\left(\frac{R\Omega}{c}\right)^2}} \tag{2-5}$$

the following equations result

$$\Delta\tau_N = \frac{4A\Omega}{c^2\left[1-\left(\frac{R\Omega}{c}\right)^2\right]} \tag{2-6}$$

$$\Delta\tau_R = \frac{4A\Omega}{c^2\sqrt{1-\left(\frac{R\Omega}{c}\right)^2}} \tag{2-7}$$

For usual angular rates in the range of

$$R\Omega \ll c$$

the selected transformation is practically not important.

Any correction terms within the range of

$$\left(\frac{R\Omega}{c}\right)^2 \Delta\tau$$

can be neglected so that the simple relation

$$\Delta\tau = \frac{4A\Omega}{c^2} \tag{2-8}$$

applies independent from the state of the observer.

After travelling the optical path both beams are superimposed to produce optical interference fringes. When the system is rotating a small shift in the position of the fringes (on the rotating platform) can be observed.

Since the measuring effect is indirectly obtained using this interferences fringe shift the following equation results for this

$$\Delta Z = \frac{4A\Omega}{c\lambda_o} \tag{2-9}$$

This interference fringe shift caused by the Sagnac-Effect was derived supposing that the light propagates in a vacuum. For light propagation in a medium (e. g. gas) basically differing velocities of light have to be considered for the two counter rotating beams.

Taking into account applications in guidance and control it can be shown however that equation (2-9) also applies for a light path within a medium /54/.

Doppler correction:

In order to be complete the doppler shift of the read-out wavelength

between the stationary observation point and the rotating interference point (beam splitter) has to be taken into account for the measured wave length.

Due to the doppler effect /2/ the following fringe shift change results for a relative velocity between beam splitter and observer:

$$\delta(\Delta Z) = \frac{4A\Omega}{c\lambda_0\left(1+\frac{v}{c}\right)} - \frac{4A\Omega}{c\lambda_0} = \left(\frac{4A\Omega}{c\lambda_0}\right)\cdot\frac{v}{c} \qquad (2\text{-}10)$$

For a circular path with the radius R = 0.1m this correction is approximately:

$$\frac{R\cdot\Omega}{c} = \frac{v}{c} = 6\cdot10^{-9} \qquad (2\text{-}11)$$

at an angular rate of 1000deg/s and can thus also be neglected for all practical applications considered here.

2.3 Resolution and Scale Factor of the Sagnac Effect

The resolution obtainable with an angular rate measurement device using the Sagnac Effect are presented in table 2-1.

For two different sized areas spanned by an assumed circular light path (0.01 m^2 and 1 m^2) and for angular rates in the range of 0.01deg/h to 10^6deg/h the table shows the non-reciprocal path differences caused by the Sagnac Effect and computed from

$$\Delta L = c\cdot\Delta\tau = \frac{4A\Omega}{c} \qquad (2\text{-}12)$$

Corresponding propagation time differences and/or fringe pattern shifts are to be resolved by the measurement equipment. To get a feeling about the order of magnitude of the path differences ΔL it shall be mentioned that the diameter of the hydrogen atom is approximately 1 Å unit (1 Å $= 10^{-8}$ cm).

As shown in the table in case of this type of measurement arrangement with dimensions suitable for inertial sensors the simple Sagnac Effect leads to unfeasible measurement resolutions in particular assuming low rates (e. g. 0.01deg/h).

Now the question arises which possibilities exist to measure the path difference ΔL as accurately as required or even possible applying technologies available today or in the near future.

The answers to this question and thus the application of the Sagnac- Effect in the realization of optical rate measurement devices have led to different designs where it is attempted to amplify the measured effect and thus increasing the resolution by

- a multi-turn of the light beam in a fibre coil
- the multiple transit of beams in only one optical ring
- the integration of a laser medium into a mirror ring arrangement causing frequency shifts of an optical resonator.

(see also fig. 2-4).

In all these concepts the application of the difference measurement principle (2 counter-rotating beams, 2 resonances for counter- rotating beams) is decisive. The above mentioned different realization concepts lead to the classification of the Sagnac Effect based optical rate sensors as shown in fig. 2-5.

3. OPTICAL RATE SENSORS, AN OVERVIEW

3.1 Introductory Remarks and Historical Experiments

In the preceding section it has been explained that all optical rate sensors are based on the Sagnac Effect.

Furthermore it has been quantitatively shown that the measurement resolution which can be achieved by the simple Sagnac Effect is too small for realizing an applicable rate sensor.

Then three different realization concepts were mentioned (see fig. 2-4) which are to lead to a considerable improvement of the measurement resolution.

Prior to describing the function principle of the mentioned concepts some historical optical experiments are to be briefly described which are of a fundamental importance for the principal function of the sensors treated here.

In the 19th century optical arrangements like that applied to explain the afore mentioned Sagnac Effect were designed in order to furnish evidence of postulats referring to the so called "ether theory".

The spectacular experiment which Michelson and Morley designed in 1887 for measuring the light propagation in a translatory moving reference system was one of these arrangements. Michelson and Morley wanted to determine the effect of the earth's movement in space on the velocity of light measured directly on its surface.

In its sun orbit the earth has a velocity of approcimately 30km/sec. According to the then prevailing physical understanding an "ether wind" was assumed to blow across the surface of the earth and perhaps even through it. This wind was expected to have the same effect on the travelling times of two beams propagating in the "ether wind" direction and across it than the current of a river has on the travelling times of a boat which moves with the current and across it.

The setup of their experiment is presented in fig. 3-1.It was mounted on a thick marble plate with a diameter of approximately 3m which floated in mercury in order to assure the performance of an undisturbed 90deg. azimuth rotation for a 2-position measurement.

From a light source (lamp) a beam hit a silver-coated glass disk P1 which let one half of the beam pass and reflected the other half. These two beams were now reflected by the mirrors M1 and M2 which were located at the same distance from the center. The beam reflected by M1 partially passed the glass disk P1 whereas the beam returning from M2 was partially reflected.

Both beams were superimposed in the telescope. The glass disk P2 was introduced in order to compensate for the additional path in P1 for the beam directed towards M2.

By making use of the translatory velocity of the earth when moving in its sun orbit and due to the resulting so-called "ether wind" which in the experiment blows e. g. from right to left, the influence of this "wind" on the propagation of light was to be investigated.

If there was no "ether wind" both beams would hit the telescope in the same phase and would cause maximum brightness. But if there was an "ether wind" the beam travelling across it would be less delayed than that beam travelling with and against it. Then a partially erasing interference would have to be expected. This was to be determined by measuring the interference fringe shift in two azimuth positions of the arrangement, set off by 90deg. It can be shown that with the translatory velocity of the earth in its orbit a 5% fringe shift would occur which was indeed measurable even with the then available devices.

Why this experiment could reveal no measurable effect - much to the disappointment of the physicists - was postulated by Einstein many years later (c = const.)

Unlike Michelson and Morley Sagnac investigated in 1911 the propagation of light in a rotating reference system. The setup of the experiment is shown in fig. 3-1 (on the right side). The mirror ring arrangement comprised an area of 866cm^2 and caused an interference fringe pattern shift of 0.07 at a rate of 2 revolutions per second.

Although the sharpness of the interference fringes was limited due to the lack of monochromatic light Sagnac could resolve the aforementioned shift.

Thus he succeeded to furnish proof of the fact that the propagation velocity of the light is not affected by the rotation of the reference system.

In 1925 Michelson and Gale repeated the Sagnac experiment in an enormously scaled-up version. On the Chicago University campus they realized an interferometer ring covering an area of 207400 m^2(340 m x 610m) which was exactly aligned to the horizon of the earth. Due to the vertical component of the earth's rotation of approximately 10deg/h at the location of the experiment a fringe shift of 0,23 was generated.

This Michelson-Gale experiment was designated the optical analogous to the Foucault pendulum (para 1.1).

In order to obtain fringe shifts that could be measured using the then available devices either gigantic measurement arrangements or high angular rates were required. The experiments revealed however that the velocity of light remains constant independent from the rotation of the reference system in which it has been measured, (see also fig. 2-2).

Some decades later the practical application of these optical rate measurement methods was again successfully envisages, which will be shown later here.

For this purposes the realization concepts indicated in fig. 2-4 will be briefly described in the following.

3.2 Passive Rate Sensors

3.2.1 Sagnac Interferometer

The Sagnac Interferometer represents the most simple realization of the Sagnac Effect described in para 2.2. It belongs to the passive interferometers using interferometeric methods for detecting the beam path differences whereas the frequency of the carrier wave is not effected by the Sagnac Effect.

The principle of such a measurement arrangement is shown in fig. 3-2. The light radiated from the source is split into two counter rotating beams along the beam splitter.

After one revolution the two beams meet again in the beam splitter, are superimposed which results in an interference fringe pattern on the image plane.

The phase shift between both beams

$$\varphi = 2\pi \frac{\Delta L}{\lambda} \tag{3-1}$$

which results from the Sagnac path difference

$$\Delta L = \frac{4A\Omega}{C} \tag{3-2}$$

leads to an intensity modulation of the interference image

$$I = I_0 (1 + \cos\varphi) \tag{3-3}$$

At a constant angular rate Ω of the arrangement this periodic fringe pattern is fixed stationary. In order to determine an angular rate of 0.01deg/h with a sensor geometry of $A = 0{,}01 m^2$, $\lambda = 0{,}6\,\mu m$ a fringe shift of

$$\Delta Z = \frac{4A\Omega}{C\lambda} = 10^{-11}$$

would have to be resolved (see also fig. 2-3 and Tab. 2-1).

The intensity to be measured at a particular point of the image plane can also be considered as a means for measuring the angular rate. For this purpose the equations (3-1) and (3-2) were implemented into equation (3-3) leading to

$$I = I_0 \left[1 + \cos\left(8\pi \cdot \frac{A}{C\lambda} \cdot \Omega \right) \right] \tag{3-4}$$

This equation (3-4) represents the relation of the intensity of the interference fringe pattern at a defined point of the image plane as a function of the angular rate with which the arrangement rotates.

It is an interesting question how such an arrangement reacts when the angular rate is changed. The differentiation of the above equation leads to:

$$\frac{dI}{d\Omega} = -I_0 \cdot \frac{8\pi A}{C\lambda} \sin\left(\frac{8\pi A}{C\lambda} \cdot \Omega \right) \tag{3-5}$$

Thus the maximum sensitivity is

$$|\Delta I|_{max} = I_0 \cdot \frac{8\pi A}{C\lambda} \Delta\Omega \tag{3-6}$$

For quantitatively showing this intensity change to be resolved the following figures are to be assumed.

$I_0 = 1\,mW$

$\lambda = 0{,}6\,\mu m$

$A = 0{,}01\ m^2$

$C = 3 \cdot 10^8\ m/sec$

$\Delta\Omega = 0{,}01\ °/h$

Thus the maximum intensity change results to be

$$|\Delta I|_{max} = 7 \cdot 10^{-14} \text{ W}$$

This means that a photodetector must be able to resolve a light intensity change of 7.10^{-14} W for a rate change of 0.01deg/h (see also fig 3-3).

This exceeds the measurment sensitivities which can be realized in practice today with affordable expense. It is the same order of magnitude with which image signals from the U.S. Mars space probe where received on earth in 1967. Sufficient intensity to recontruct almost perfect images - however, requiring instrumentations unaffordable for a simple rate sensor.

3.2.2 Multi-Turn Sagnac Interferometer (Fiber Optic Gyro)

The availability of low-loss light-conductive fibres allows to significantly increase the scale-factor by realizing a multiple turn fibre coil.

This arrangement is presented in fig. 3-4 as Multi-Turn Sagnac Interferometer. In case of N coil windings the fringe shift is

$$\Delta Z_N = N \cdot \frac{4A}{C\lambda} \cdot \Omega \qquad (3\text{-}7)$$

and the associated phase shift

$$\varphi = N \cdot \frac{8\pi \cdot A}{C \cdot \lambda} \cdot \Omega \qquad (3\text{-}8)$$

By implementing equation (3-8) into equation (3-3) the following results for the intensity of the fringe pattern at a certain point on the image plane

$$I = I_0 \left[1 + \cos \left(N \cdot \frac{8\pi A}{C\lambda} \cdot \Omega \right) \right] \qquad (3\text{-}9)$$

Compared to the single-turn Sagnac Interferometer the number of interference fringes increases with the number of windings and the angular rate Ω. The same intensity function results as shown in fig. 3-2 if the abscissa is scaled with Ω/N instead of Ω.

In order to be able to estimate the reaction of such an arrangement when the angular rate Ω is changed the equation (3-9) is differenciated.

$$\frac{dI}{d\Omega} = -I_0 \, N \cdot \frac{8\pi A}{C\lambda} \cdot \sin \left(N \cdot \frac{8\pi A}{C\lambda} \cdot \Omega \right) \qquad (3\text{-}10)$$

The following maximum sensitivity results:

$$|\Delta I|_{max} = I_0 \, N \cdot \frac{8\pi A}{C\lambda} \cdot \Delta\Omega \qquad (3\text{-}11)$$

In order to quantitatively clarify the measuring sensitivity according to equation (3-11) the previously used figures are employed here for $N = 10^3$ too (corresponds to a fibre length of approximately 400m).

In case of an angular rate change of $\Delta\Omega$ = 0.01deg/h the corresponding maximum intensity change $|\Delta I|_{max} = 7.10^{-11}$ W results. As expected the sensitivity has improved by the factor $N = 10^3$ (number of windings) which ofcourse is more favourable than in case of the simple Sagnac interferometer. A photodetector located in the image plane for detection should thus be able to resolve a light intensity change of 7.10^{-11} W for a rate change of 0.01deg/h. The characteristics is ambigous with respect to the angular rate.

For achieving the maximum sensitivity there are several possibilities of how to introduce a bias the phase effect of which will be that the operating point can be moved to where the intensity characteristic curve has it maximum slope (e. g. phase shifter in one of the two light beams).

How the fiber optic gyro is derived from the basic Sagnac Effect is summarized in fig. 3-5.

3.2.3 Passive Ring Resonator (PARR)

Instead of using a multi-turn arround the interferometer area (as in the Multiple-Turn Sagnac Interferometer) the measurement sensitivity is here increased by a multiple transit of a light beam in only one optical ring.

The function principle of the instrument is presented in fig. 3-6 and represents a Fabry-Perot Interferometer.

An external light source (laser, monochromatic, wave length λ) introduces a beam through the semi-transparent mirror I into the ring interferometer consisting of mirrors I, II and III. By this optical mirror arrangement the entered beam is trapped for multiple transits within the closed optical path.

The beam passes the given path an infinite number of times: after each circulation a fraction of the beam intensity is output at mirror II. Between the mirrors I, II, III the waves travel on absolutely identical optical paths. In order to illustrate the multiple transit effect different optical paths are shown in fig. 3-6.

The individual waves E_1 , E_2 .... E_n that are decoupled and interfered at mirror II have a mutual phase difference of

$$\varphi = \frac{2\pi}{\lambda} \cdot (L + \Delta L) \qquad (3\text{-}12)$$

where ΔL represents the path difference caused by the known Sagnac Effect. When A_o is the amplitude of the primarily introduced wave this yields for the amplitude of the electric field at the output:

$$E_p = A_o \exp\left[j 2\pi \nu \left(t - \frac{L}{C}\right)\right] \cdot T \cdot \left[1 + Re^{-j\varphi} + R^2 e^{-j2\varphi} + \ldots\ldots R^{P-1} e^{-j(P-1)\varphi}\right] \qquad (3\text{-}13)$$

T and R represent the transparency and reflectivity factors of the light amplitudes (T = 1-R).

The geometrical progression can be presented in the closed form for $p \longrightarrow \infty$

$$E_\infty = A_o \cdot T \cdot \exp\left[j 2\pi \nu \left(t - \frac{L}{C}\right)\right] \frac{1}{1 - Re^{-j\varphi}} \qquad (3\text{-}14)$$

The multiplication by the conjugate complex value yields the intensity of the resulting wave

$$I_\infty = I_o \cdot \frac{T^2}{(1-R)^2 + 4R \sin^2\left[\frac{\pi}{\lambda}\left(L + \frac{4A}{C} \cdot \Omega\right)\right]} \qquad (3\text{-}15)$$

Again a relation exists between the intensity of the superimposed beam fractions decupled at mirror II and the angular rate Ω with which the arrangement rotates about an axis perpendicular to the area spaned by the closed optical path.

In fig. 3-6 the intensity distribution at a particular point of the image plane is shown as a function of the mutual phase difference of decoupled beams which depends on the angular rate Ω .

If the perimater L (optical path) is made an integer multiple of the wave lenght of the introduced laser light ($L = n.\lambda$) the arrangement becomes a passive ring-resonator with the resonance frequency $f_o = c/\lambda$.

The "sharpness" of the interference fringes depends on the quality of the resonance in the ring resonator. The lower the losses at the mirrors (R high) the "sharper" will be the resonance and the steeper the interference fringes which signifies a high measurement sensitivity.

This behaviour can be compared to a mechanical resonator which shows a considerable resonance amplitude margin in case of low damping.

The quantitative sensitivity of the arrangement results from the differentiation of equation (3-15) with respect to Ω

$$\frac{dI_\infty}{d\Omega} = - \frac{I_o T^2 4R \sin\left[\frac{2\pi}{\lambda}\left(L + \frac{4A}{C} \cdot \Omega\right)\right] \cdot \frac{4\pi A}{\lambda C}}{\left\langle (1-R)^2 + 4R \sin^2\left[\frac{\pi}{\lambda}\left(L + \frac{4A}{C} \cdot \Omega\right)\right]\right\rangle^2} \qquad (3\text{-}16)$$

Using the figures of para 3.2.1 the following highest possible light intensity change per angular rate change $\Delta\Omega$ is obtained

$$|\Delta I_\infty|_{max} = 0{,}45 \cdot 10^{4} \Delta\Omega$$
$$|\Delta I_\infty|_{max} = 2{,}1 \cdot 10^{-12}\ W \qquad (3\text{-}17)$$

Fig. 3-7 shows two principles for signal detection procedures. The analog intensity measurement is based on the phase difference between the individual waves indicated in equation (3-12). The resonance tuning $L = n \cdot \lambda$ yields for $\Omega = 0$ a maximum intensity with $\varphi_0 = 2\pi n$ which would be measured by the detector.

If the arrangement rotates ($\Omega \neq 0$) the intensity change would be minimal because of the horizontal tangent of the intensity distribution within the interferences fringe around φ_0.

In order to make use of the maximum sensitivity a bias phase shift $\Delta\varphi_B$ is introduced, e. g. by a small change of the resonator length. By this the operating point is shifted accordingly on the intensity distribution curve to maximize the sensitivity for changes in phase difference (i. e. $\Delta\Omega$). Around this operating point the intensity is proportional Ω and can thus be used as an analog measurement signal for the angular rate.

As already mentioned the slope of the intensity distribution curve depends on the resonator quality. The second method for signal extraction which is basically feasible is the indirect frequency measurement also shown in fig. 3-7.

Due to the aforementioned tuning $L=n\cdot\lambda$ the arrangement is in resonance for the frequency $f_0 = c/\lambda$ at $\Omega = 0$. The detector measures the corresponding maximum intensity.

An angular rate of $\Omega \neq 0$ detunes the ring resonator since the effective resonator length changes so that the arrangement is no longer in resonance for the frequency introduced by the external lightsource. Thus the detector measures a lower intensity.

A servo electronics adjusts a "frequency shifter" in accordance with the measured intensity difference. The frequency shifter is located in the beam path between the external light source and the resonator and continues to change the frequency of the light introduced into the ring until the intensity measured by the detector has reached the maximum value again.

The resonator is in resonance again and the set frequency is $f = f_0 + \Delta f$. The frequency shift Δf corresponds to the one which occurs in the active ring laser gyro, too (see also equation (3-20) and is proportional Ω. It can be measured indirectly using the signal which is fed into the frequency shifter (see also fig. 3-7).

The measurement effect may be further increased by using two counter rotating wave beams in the mirror ring arrangement, same as in the Sagnac Interferometer. During the rotation of the arrangement these beams show opposite phase shifts at the output mirror. The order of magnitude of the previous assessments remains however the same.

Fig. 3-8 summarizes how the PARR gyro is derived from the basic Sagnac Effect.

The interferometer type treated here can be realized by using a mirror arrangement as mentioned, glass fibre or wave guides.

3.3 The Active Ring Resonator

3.3.1 Introductory Remarks

The rate sensors described so far in para 3.2 are so called passive instruments.

They are characterized by the feature that an external light emitter (e. g. laser) radiats into an interferometer or ring resonator. Depending on the angular rate of the arrangement the measurable light intensity changes; this change represents the angular rate and is determined basically in an analogue type measurement.

The quantitative assessments have revealed that in case of angular rates down to 0,01deg/h extremely small intensity changes have to be detected and resolved.

Unlike the passive sensors the ring laser gyro is an active sensor where a laser medium is integrated into a mirror ring arrangement (active ring resonator). Here the laser medium acts as an amplifier (active element) for the light with a particularly wave length.

Prior to dealing with the derivation of the ring laser gyro (RLG) function from the Sagnac Effect some remarks are to be made concerning the gas laser.

3.3.2 Brief Description of the Gas Laser Mechanism

The term "Laser" is an abbreviation for "Light Amplification by Stimulated Emission of Radiation".

As far as rate sensors are concerned gas lasers (mostly helium-neon) are particularly suitable for permanent operation. They require few power and are almost monochromatic.

An amplifying medium and a resonator cavity are prerequisits in order to achieve the laser effect.

The amplifying medium consists of a closed tube containing a low- pressure mixture of e. g. helium and neon gas (fig. 3-9).

The gases are ionized by a voltage which is applied to a metal anode and cathode, igniting a glow discharge.

During this process excited helium atoms collide with neon atoms thus transferring energy to them so that they reach an excited energy state (pumping).

Within a very short time the neon atoms fall back into their original energy level by spontaneously emitting one photon (fig. 3-9).

Emitted photons travelling through the medium collide with other excited neon atoms which then emit photons, too (stimulated emission). Repeating these processes causes a progressive photon emission in all directions. If the gas discharge tube is closed on both ends using mirrors there will be a cavity resonator in which the photons which travel parallel to the tube axis are reflected to and between the two mirrors. This cavity resonator furnishes the optical feedback which is required for an undamped oscillation.

The cavity is in resonance for those frequencies the wave lengthes of which are integer dividers of the resonator length. This applies for many frequencies since the length of the resonator is many thousand times longer than the optical wave length (see also fig. 3-10). The majority of these frequencies are not adequately amplified and thus they do not result in oscillations.

An undampened oscillation occurs only for those resonance frequencies whose amplification exceeds the losses.

Usually approximately 0,2% of the radiation intensity of the light beam is coupled out via the end mirror of the laser tube. This laser radiation essentially consists of only one frequency (monochromatic radiation). The frequency of the laser radiation depends on the length of the cavity resonator.

3.3.3 The Ring Laser Gyro (RLG)

A gas laser tube is integrated into a triangle mirror ring arrangement as shown in fig. 3-11. Again the laser tube acts as an amplifier of light with a particular wave length. The mirrors are used for the feedback of this optical triangle resonator. Now the radiation is no longer reflected between to mirrors but passes a closed optical path in both directions, where one beam travels cw (a-beam) the other ccw (b-beam). Both beams are at the same time in the resonator cavity and have the same frequency in case of a stationary arrangement. If the arrangement is rotated at the angular rate Ω one beam has to travel a longer distance and the other one a shorter distance for one circulation.

Due to the Sagnac Effect a path difference of

$$\Delta L = \frac{4A}{C} \cdot \Omega \qquad (3\text{-}18)$$

results between the two counter-rotating waves (see also equation 2-12).

If the effective resonance cavity length L within the ringresonator changes by ΔL (see also fig. 3-11) the wave length λ_0 of the generated radiation changes by $\Delta\lambda$ or the corresponding frequency f_0 changes by Δf since the ring laser remains in resonance, however, for a different resonance frequency.

$$\frac{\Delta f}{f_0} = \frac{\Delta \lambda}{\lambda_0} = \frac{\Delta L}{L} \rightarrow \Delta f = f_0 \cdot \frac{\Delta L}{L} \qquad (3\text{-}19)$$

For a He-Ne laser f_0 is approximately 10^{14} Hz. From the preceeding equation it

becomes evident that due to the ring resonator the original Sagnac Effect is amplified by this factor thus improving the measurement resolution (equation 3-19) enormously.

When equation (3-18) is applied in equation (3-19) the following results for the frequency difference to be measured

$$\Delta f = \frac{4A}{L\lambda} \cdot \Omega \qquad (3\text{-}20)$$

Using the figures of para 3.2.1 a frequency difference of Δf =0.005 Hz results for Ω = 0.01deg/h and of Δf = 1800 Hz for Ω = 1deg/sec.

For measuring the frequency difference the two counter-rotating waves are decoupled at the output mirror (see also figure 3-11 and 3-12). By superimposing these two signals of different frequencies an interference pattern results which has the character of a beat the frequency of which corresponds to the difference frequency (equation 3-20).

As shown in fig. 3-12 the beat moves across a photo detector in the image plane. The interference fringes are counted when passing the photo detector; the number of fringes passing per unit of time represents a measurement for the angle change ($\Delta\Theta$) caused by Ω during the corres ponding period of time.

From equation (3-20) and with N being

$$N = \int_0^t \Delta f \, dt \qquad (3\text{-}21)$$

the number of fringes per time period as well as

$$\Delta\Theta = \int_0^t \Omega \, dt \qquad (3\text{-}22)$$

being the angular change per time period the following results as relation between angular change and number of passing fringes:

$$N = \frac{4A}{L\lambda_o} \cdot \Delta\Theta \qquad (3\text{-}23)$$

Again using the figures of para 3.2.1 euqation (3-23) leads to $N \approx 10^5 \cdot \Delta\Theta$. Thus one passing interference fringe corresponds to an angular change of 2 arcsec which is a very good resolution.

Unlike the previously treated passive optical rate sensors the ring laser gyro furnishes a digital output signal. With the type of the indicated signal extraction it possesses the feature of a "rate integrating gyro".

Fig. 3-13 summarizes the derivation of the ring laser gyro from the basic Sagnac Effect.

4. FUNDAMENTAL PHYSICAL LIMITATIONS OF OPTICAL RATE SENSORS

So far three different concepts for optical rate sensors based on the Sagnac-Effect have been trated as far as their fundamental functions are concerned.

It has been shown that these concepts offer sufficient measurement resolution and thus they represent a basis for respective sensor developments.

Finally the fundamental physical (not man made) limitations shall be considered here, where the results of which are summarized in fig. 4-1.

As far as the multi-turn Sagnac Interferometer is concerned the photon shot noise is the limiting factor resulting in a noisy intensity pattern as shown in fig. 3-2.

As can be seen from the approximate equation in fig. 4-1 the angular rate sensing limit can be reduced by increasing the number of fibre turns, however this increases the losses in the fibre.

Since the source is a stochatic process its effective quantity depends on the measurement interval (T) available for averaging the sensor output signal.

The photon shot noise is again the limiting source for the passive ringresonator resulting in an approximate formula again shown in fig. 4-1. From this it can be seen, that the limit in angular rate sensing capability can be reduced by increasing the quality of the passive optical resonator; that is reducing Γ, which corresponds to the bandwidth of the resonance peak as shown in fig. 3-7.

Regarding the active ringresonator the frequency difference of two optical resonators is a measure for angular rate. Phase fluctuations caused by spontaneous emission of photons are the fundamental limit to sense angular rates with the active ringresonator. They cause stochastic fluctuation of the resonance frequencies f_a and f_b of the counter rotating beams (see fig. 3-11). Again an approximate formula is given in fig. 4-1 which is similar to that shown for the passive ringresonator.

For the example data applied as representative numbers in fig.4-1 nearly identical results are obtained for all three sensor-types. The applied averaging time affects the results for the three sensors in exactly the same way.

As far as the active ringresonator is concerned, the result mentioned in fig. 4-1 has already been confirmed by existing instruments.

Regarding the passive sensors a breakthrough towards quantities as mentioned in fig. 4-1 has not been demonstrated so far. There are however fundamental possibilities to even arrive at results beyond these numbers by increasing the length of the fiber within the multiturn interferometer or by increasing the resonator quality (reduce Γ) as far as the passive resonator is concerned.

It has to be taken into account however that the fundamental limit of the active ringresonator can also be decreased further if the path length and quality are increased.

For all three sensor types the fundamental limits obtained go beyond that order of magnitude which must be met in order to built inertial systems with a performance of better than 0,5 NM/h.

5. CONCLUDING REMARKS

It has been shown that there is a wide range of guidance and control applications for which rate sensors with broadly differing requirements are needed.

After dealing with the reasons of optical rate sensor penetration in the near future a short introduction to the physical function and basic design of passive and active optical rate sensors has been given. The main statements are summarized again in fig. 5-1.

Particularly the passive sensors are expected to yield a decisive breakthrough in the future since they seem to offer a more favourable cost versus performance potential as compared to the active ring laser gyro.

The realization of the relatively simple physical principles of optical rate measurements as treated here involves new gyro technology problems which now refer to the electro-optics, laser physics and integrated optics rather than to the precision mechanics.

It is expected that feasibility problems still existing today regarding passive sensors will be alleviated in the future where the inertial field can benefit from developments within the optical communication field.

As far as the active ring laser gyro is concerned it has obtained a high level of maturity being ready for mass production. However, it is still an expensive sensor.

The individual problems which will be encountered for the realization of the three sensor types described here are treated in more detail in /52/, /53/ and /54/.

Literature

/1/ Sagnac, G.: Compt.Rend.157, 1913, pp 708, 1410
J.Phys.Radium, 5th Ser., 4. 1914. pp 117.

/2/ Post, E.J.: Rev. of mod. Phys. Vol 39, No.2, April 1967, pp 475

/3/ Heer, C.V.: The Physical Review, Vol. 134, No. 4A, May 1964, pp A 799

/4/ Leeb, W.R.: Schiffner, G.; Scheiterer, E.:
Applied optics, Vol. 18, No.9, May 1979, S. 1293-1295

/5/ Kromykh, A.T.:Soviet Physics-JETP 23, 1966, pp 185

/6/ Kleen, W.; Müller, R.(Hrsg):
Laser Springer-V. Berlin, 1969

/7/ Röss, D.: Laser Lichtverstärker und Oszillatoren, Akad. Verlagsgemeinschaft
Frankfurt/M., 1966

/8/ Gürs, K.: Laser, Umschau-Verlag, Frankfurt/M. 197o

/9/ Aronowitz,F.: The Laser Gyro, in Monte Ross (ed.):
Laser Applications Vol. I, Acad.Press, New York, 1971

/10/ Lamb, W.E.: Physical Review, Vol. 134 No. 6A, 1964, pp A 1429

/11/ Aronowitz,F.: Lim, W.L.: IEEE Journal of Quantum Electronics
Vol. QE-13, No.5, May 1977, pp 338

/12/ Rodloff, R.: Symp. Gyro Technology, Stuttgart, 24./25.Sept. 1980

/13/ Aronowitz,F.: Collins, R.J.: Journ. of Appl. Phys. Vol. 41, No.1, Jan.197O, pp 130

/14/ Podgorski,T.J.: Aronowitz, F.:
IEEE Journal of Quantum Electronics, Vol. QE-4, NO.1, Jan.1968, pp 11

/15/ Lee, P.H.;Atwood, J.G.:
IEEE Journ.of Quantum electronics, Vol.QE-2, No.8, Aug.1966, pp 235

/16/ Hutchings, T.J. et.al: Phys.Review, Vol. 152, No.1, 2 Dec. 1966, pp 467

/17/ Papoulis, A: Probability, Random Variables and Stochastic Processes.
Mc.Graw Hill, 1965

/18/ Anderson, L.K.; Mc.Murtry, B.J.
Appl.Opt./Vol. 5, No.10/1966 pp. 1573

/19/ Melchior,H, et.a.:
Proc.IEEE, Vol 58, No.10, Oct.197O, pp 1466

/20/ Shih-Chun Lin.; Giallorenzi, T.G.:
Applied Optics, Vol 18, No.6/15 March 1979/ pp 915

/21/ Hutchings, T.J.; Stjern, D.C. (Litton Industries):
Proc.NAECON 1978, 549-55, N.Y., IEEE

/22/ Chow,W.W.; et.al:
IEEE Journ. of Quantum Electronics, Vol QE-14, No.9, Sept.198O, pp 918

/23/ Dorschner, T.A. et al.:
Proc.NAECON, 1978, 569-573, N.Y. IEEE

/24/ Mc.Clure, R.E; Vaher, E.:
Proc.NAECON, 1978, 544-548, N.Y. IEEE

/25/ Thomson,K.: Proc.SPIE, Vol 157, pp 13-2O, 1978

/26/ Scully, M.O.: Optics letters, Vol 3, No.2, pp 43, 1978

/27/ Aronowitz,F.; Lim,W.L.:
Opt. Eng. Vol 18,4,1975

/28/ Boot, H.J. et al.:
Electron.Lett. Vol 5, pp 347-348, July 1969

/29/ Coccoli, J.D.; Lawson, J.R.:
U.S.Patent 3533 O14, Oct. 197O

/30/ Schlemper, E; Mohr, F.:
CS/ EROA, Okt. 198O, SEL

/31/ Cahill, R.F.; Udd, E.:
Optics Letters, Vol 4, No.3, March 1979, pp 93

/32/ Thomson, D.E., et.al.:
Appl. Phys. Lett., 33 (11; 1 Dec. 1978, pp 940)

/33/ Schiffner, G.: Dt.Pat.Nr. 2936 248, 1979, Siemens AG

/34/ Böhm, M.: Dt.Pat.Nr. DE 2917399, 1979, SEL

/35/ Ezekiel, S.; Balsams, S.R.:
Applied Phys. Lett. Vol 30, No.9, 1 May 1977, pp 478

/36/ Ezekiel, S.; et.al.:
Proc.SPIEL, Vol 157, pp 68, 1978

/37/ Hunter, J.S.; Stripling, W.W.:
Sympos. Gyro Techn. 1980, DGON

/38/ Kayton, M.: Fundamental Limitations of Inertial Measurements, Guidance and Control, Vol.8, Academic Press, New York 1962

/39/ Katt, D.R.: The Physics of Inertial Measurement Limitations, J.S.D.E.-Conference-Proceedings, 1980

/40/ Knausenberger, J.E.:
Optical Rotation Rate Sensors, AGARDograph No. 254, 1981

/50/ Rodloff, R.: Laserkreisel und andere optische Drehgeschwindigkeitssensoren DFVLR Braunschweig, 1979

/51/ Bernard, W.: Optische Sensoren für Drehgeschwindigkeit und Beschleunigung Interner Bericht, Bodenseewerk, 1981

/52/ Aronowitz, F.: Overview of physical, theoretical background of the Ring-Laser Gyro, status and problem areas.
AGARD Short Course "Optical Rate Sensors", Stuttgart and London, Sept. 1982

/53/ de Salabery, B.:
Selection of Ring Laser Gyro technologies with respect to specific applications
AGARD Short Course "Optical Rate Sensors"
Stuttgart and London, Sept. 1982

/54/ Ezekiel, S.: Passive Optical Gyroscopes
AGARD Short Course "Optical Rate Sensors"
Stuttgart and London, Sept. 1982

/55/ Schulz-Du Bois, E.O.
Alternative Interpretation of Rotation Rate Sensing by Ring Laser
I.E.E.E. Journal of Quantum Electronics Aug. 1966

/56/ Sperry Engineering Review, Elektro-Optics
Sperry Rand Cooperation, Great Neck, N.Y. 1966

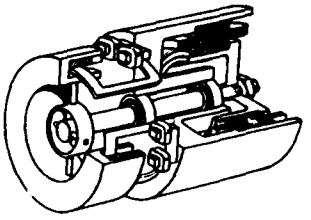
/57/ Gamov, G.: Biographie der Physik
ECON-Verlag, Düsseldorf, 1965

PHYSICS:

- ANY ROTATING BODY

ENGINEERING:

- FAST SPINNING ROTOR
- COMPLETE UNIT CONSISTING OF
  - o ROTOR
  - o DRIVING MOTOR
  - o GIMBALLED SUPPORT
  - o ELECTRONICS



INERTIAL TECHNIQUES:

- UNIT MEASURING ANGULAR RATE OR ANGULAR CHANGES
  - o CONVENTIONAL ELECTRO-MECHANIC GYROS (ROTOR)
  - o ADVANCED RATE SENSORS (NO ROTOR) LIKE OPTICAL RATE SENSORS

Fig.1-1: THE NOTION "GYRO"

- FIXED STARS ARE SPATIAL REFERENCE (INERTIAL REF.) FOR DEFINING DIRECTION OR DIRECTIONAL CHANGES

- METHODS FOR ANGULAR ROTATION SENSING:

  - o CELESTIAL NAVIGATION/GUIDANCE
    STELLAR OBSERVATIONS BY STAR-TRACKERS

  - o TERESTRIAL NAVIGATION/GUIDANCE

    REQUIRES SENSING OF ROTATION WITHOUT NEED FOR EXTERNAL COMMUNICATION

    - SPIN-TYPE GYROSCOPES

      ELECTROMECHANICAL (ROTOR)

      NUCLEAR (SPIN OF NUCLEI)

    - SAGNAC-EFFECT TYPE GYROSCOPES

      PASSIVE OPTICAL GYROS

      ACTIVE OPTICAL GYROS

Fig.1-2: SENSING INERTIAL ROTATION

- 1851, FOUCAULT DEVISED A VERY LONG (APROX. 70 M) WIRE PENDULUM CARRYING A HEAVY MASS, HUNG IN PANTHEON.

  OBJECTIVE: MEASURE EARTH'S RATE

- 1852, FOUCAULT DEVISED THE FIRST SPINNIG ROTOR GYROSCOPE

  OBJECTIVE: INDICATE EARTH'S ROTATION

- GYROCOMPASS, FIRST PRACTICAL GYROSCOPIC INSTRUMENT DESIGNED INDEPENDANTLY BY HERMAN ANSCHÜTZ AND ELMER SPERRY MORE THAN HALF CENTURY AGO.

  OBJECTIVE: HEADING REFERENCE

- FIRST INERTIAL NAVIGATION SYSTEM IMMEDIATELY AFTER WORLD WAR II

Fig.1-3: DEVELOPMENT OF MECHANICAL RATE-SENSORS

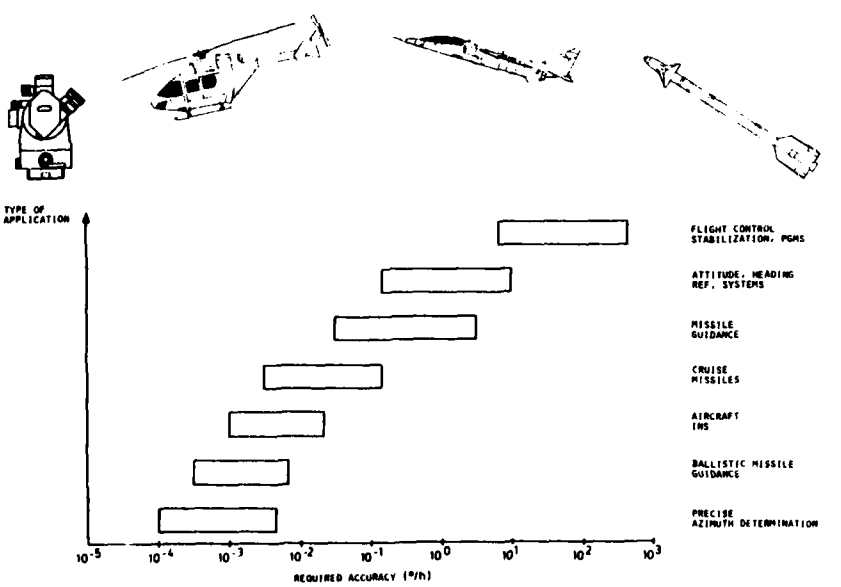


Fig.1-4: RATE-SENSOR APPLICATION VS.PERFORMANCE

- BROAD SPECTRUM OF APPLICATIONS
- LARGE RANGE OF REQUIREMENTS
  HAS LED TO THE DEVELOPMENT AND APPLICATION OF VERY DIFFERENT RATE SENSOR TYPES
  - o SIMPLE MECHANICALLY CAGED RATE-GYRO
  - o MAGNETOHYDRODYNAMIC RATE-SENSOR
  - o FLUID-PIEZOELECTRIC DUAL AXIS RATE TRANSDUCER (DART)
  - o VIBRATING WIRE RATE SENSOR (VWRS)
  - o SINGLE- AND TWO-DEGREE OF FREEDOM FLOATED GYRO (SDFG, TDFG)
  - o TWO-AXIS DYNAMICALLY TUNED GYRO (DTG)
  - o ELECTRO-STATIC SUPPORTED GYRO (ESG)
    ETC.
- NEW DEVELOPMENTS
  - o OPTICAL RATE SENSOR (ACTIVE, PASSIVE)
  - o NUCLEAR MAGNETIC RESONANCE GYRO (NMR)

Fig.1-5: TYPES OF RATE-SENSORS

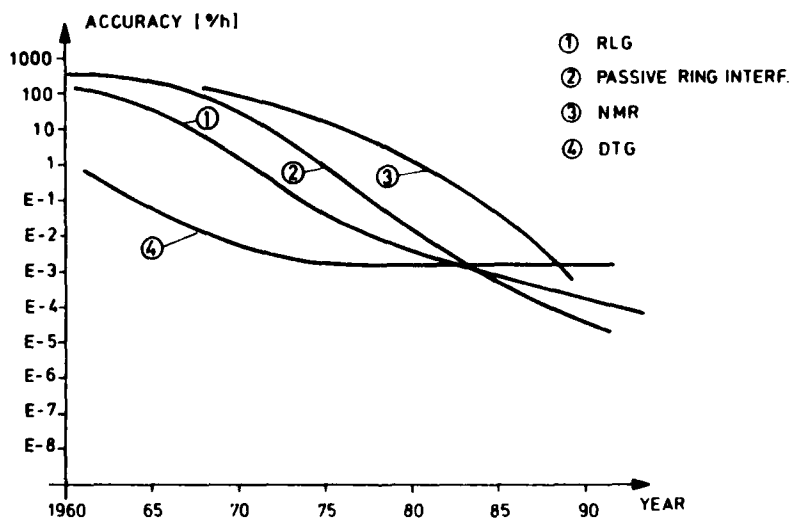


Fig.1-6: PROJECTED RATE SENSOR POTENTIAL

PHENOMENA THAT FUNDAMENTALLY LIMIT THE ACCURACY OF INERTIAL RATE-SENSING ON EARTH

- ESTABLISHMENT OF NON-ROTATING INERTIAL REFERENCE FRAME IS LIMITED BY MOTION OF THE STARS (α-CENTAURI 18μrad /YEAR)

  APROX. 10^{-7} °/h

- PRECESSION OF EARTH'S ROTATIONAL RATE VECTOR WITH RESPECT TO INERTIAL SPACE

  APROX. 10^{-5} °/h

- ROTATIONAL RELATIVISTIC EFFECT (SCHIFF PRECESSION ET.AL.)

  APROX. 10^{-8} °/h

- QUANTUM MECHANICS UNCERTAINTIES

  APROX. 10^{-20} °/h

Fig.1-7: FUNDAMENTAL RATE-SENSING LIMITS

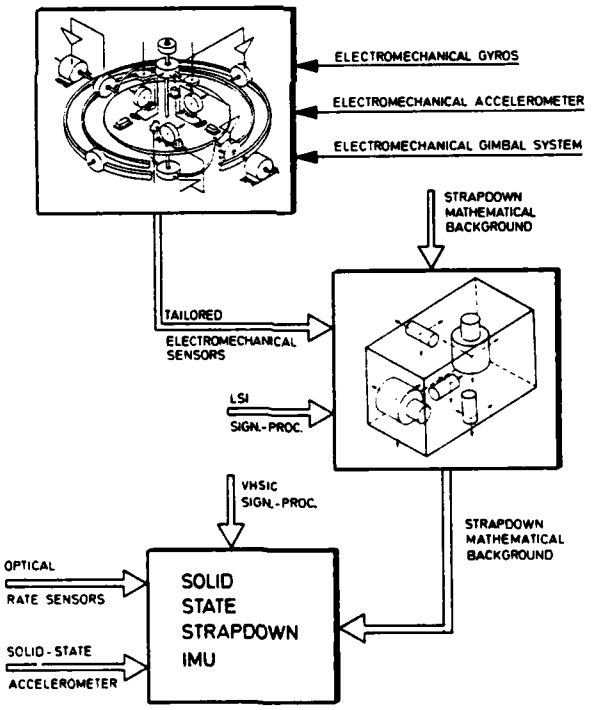
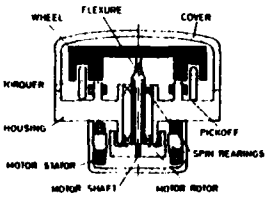


Fig.1-8: INERTIAL SYSTEM EVOLUTION

- HAVE NO MECHANICAL MOVING PARTS. THEREFORE THEY ENHANCE SIGNIFICANTLY THE SYSTEM RELIABILITY
- HAVE A WIDE DYNAMIC RANGE(800 DEG/SEC OR EVEN HIGHER)
- THEIR OUTPUT IS IN DIGITAL FORM, RESULTING IN MUCH SIMPLER INTERFACE WITH MICROPROCESSORS AND AIRBORNE COMPUTERS
- THEY ARE INSENSITIVE TO THERMAL GRADIENTS
- THEY HAVE HEATERLESS OPERATION
- INSTANT TURN-ON NEED NO TEMPERATURE CONTROL
- HAVE NO G-SENSITIVE DRIFT ERRORS
- OFFER LOW COST
- REDUCE WEIGHT AND VOLUME

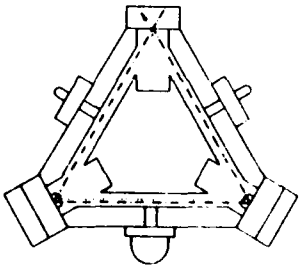
Fig.1-9: RINGLASER-GYRO ATTRIBUTES

DRY-TUNED-GYRO



Y-AXIS	X-AXIS	MEASURED ERROR-TERMS
$\frac{M_z}{H}\cdot(1+DSF_y)$	$-\frac{M_y}{H}\cdot(1+DSF_x)$	SCALEFACTOR
$\cdot[\omega_y$	$\cdot[\omega_x$	DESIRED RATE TERM
$+a_{yx}\cdot\omega_z-a_{yz}\cdot\omega_x$	$+a_{xz}\cdot\omega_y-a_{xy}\cdot\omega_z$	MISALIGNMENT
$-m\cdot a_y+q\cdot a_x+n\cdot a_y\cdot a_z$	$-m\cdot a_x-q\cdot a_y+n\cdot a_x\cdot a_z$	ACCEL DEP TERMS
$+\frac{C-A}{H}\cdot\omega_z\cdot\omega_y+\frac{A}{H}\cdot\dot{\omega}_z-\frac{E}{H}\cdot\dot{a}_i\cdot\omega_y$	$+\frac{C-A}{H}\cdot\omega_z\cdot\omega_x-\frac{A}{H}\cdot\omega_y-\frac{E}{H}\cdot\dot{a}_i\cdot\omega_x$	DYNAMIC TERMS
$+b_y]$	$+b_x]$	BIAS

RINGLASER-GYRO



X-AXIS	MEASURED TERMS
$\omega_0\cdot(1+DSF_x)$	SCALEFACTOR
$[\omega_x\cdot$	DESIRED RATE TERM
$+a_{xz}\cdot\omega_y-a_{xy}\cdot\omega_z$	MISALIGNMENT
	ACCEL DEP TERMS
	DYNAMIC TERMS
b_x	BIAS

Fig.1-10: MEASUREMENT-MODELS OF "DTG" AND "RLG"

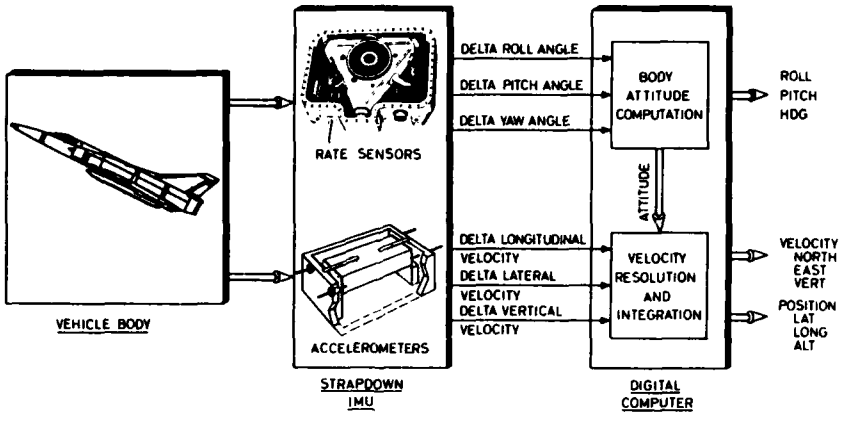


Fig.1-11: SIMPLIFIED STRAPDOWN INERTIAL SYSTEM BLOCK DIAGRAM

STRAPDOWN INERTIAL SYSTEMS UTILIZING RING LASER GYRO TECHNOLOGY OFFER SIGNIFICANT BENEFITS TO THE USER. AMONG THESE ARE:

- o SYSTEM CHARACTERISTICS
  - HIGH PERFORMANCE
  - HIGH RELIABILITY
  - LOW LIFE-CYCLE-COST
  - ELIMINATION OF ADDITIONAL SENSORS SINCE IT PROVIDES DIRECTLY BODY-AXIS INFORMATION (E.G.,RATES, ATTITUDE & HEADING)
- o MAINTAINABILITY
  - MINIMIZES MAINTENANCE ACTIONS
  - GYROS NEED NO FIELD RECALIBRATION
- o STRAPDOWN INS RESP. AHRS APPLICATIONS
  - POSITIONING/NAVIGATION
  - FLIGHT & FIRE CONTROL
  - SURVEILLANCE AND RECONNAISSANCE

Fig.1-12: STRAPDOWN LASER GYRO SYSTEMS BENEFITS

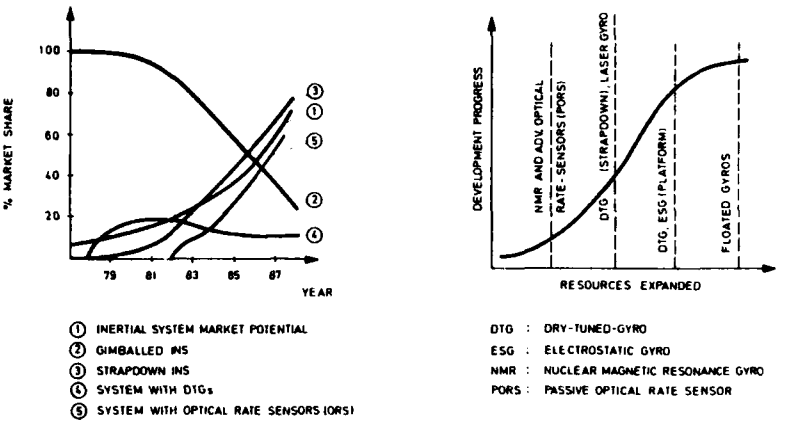


Fig.1-13: INS MARKET POTENTIAL AND ORS PENETRATION

- GROWING ROLE FOR INERTIAL TECHNOLOGY AND GROWING INS MARKET
- GROWING PORTION OF INS IN STRAPDOWN CONFIGURATION
- MAJOR COMPETITION IN THE 1980-90 PERIOD EXPECTED BETWEEN ACTIVE RING LASER GYROS AND DYNAMICALLY TUNED 2-AXIS GYROS (DTG)
- DTG'S WILL BE COMPETETIVE FOR MANY YEARS IN FUTURE AND MAINTAIN A FAIR MARKET SHARE
- GROWING APPLICATION OF OPTICAL RATE SENSORS IS EXPECTED PROPORTIONAL TO THE INCREASING SHARE OF STRAPDOWN SYSTEMS
- A BREAKTHROUGH IN THE DEVELOPMENT OF SOME OF THE PASSIVE OPTICAL RATE SENSORS (LOW COST) COULD SHIFT THE MARKET DRAMATICALLY TOWARDS SUCH SENSORS

Fig.1-14: CONCLUSION INS MARKET POTENTIAL AND ORS PENETRATION

PARAMETER	DYNAMICALLY TUNED GYRO (DTG)	RING-LASER GYRO (RLG)
COST	?	?
RELIABILITY	SAME	SAME
SIZE	SMALL	LARGER
PERFORMANCE		
RATE	FAIR	GOOD
BIAS	GOOD	GOOD
RANDOM NOISE	GOOD	FAIR
SCALE FACTOR	FAIR	GOOD
REACTION TIME	FAIR	GOOD
MATURITY	GOOD	FAIR

Fig.1-15: COMPARISON OF STRAPDOWN GYRO-TYPES

- THERE IS NO DIFFERENCE IN KIND BETWEEN THE TYPE OF INERTIA POSSESSED BY MECHANICAL MASS AND THAT POSSESSED BY ELECTRO-MAGNETIC RADIATION.

- ROTATION SENSING IS A MANIFESTATION OF THE INERTIAL CHARACTER OF ELECTROMAGNETIC RADIATION.

- THE RING-RESONATOR WITH ITS STORED ENERGY IS ANALOGOUS TO A HEAVY RING-SHAPED MASS CONTAINED INSIDE A FRICTION FREE BEARING SURFACE.

- SCHULZ - DU BOIS :

  ALTERNATIVE INTERPRETATION OF ROTATION RATE SENSING BY RING LASER [55]

Fig.2-1: MASS-ENERGY-EQUIVALENCE [55]

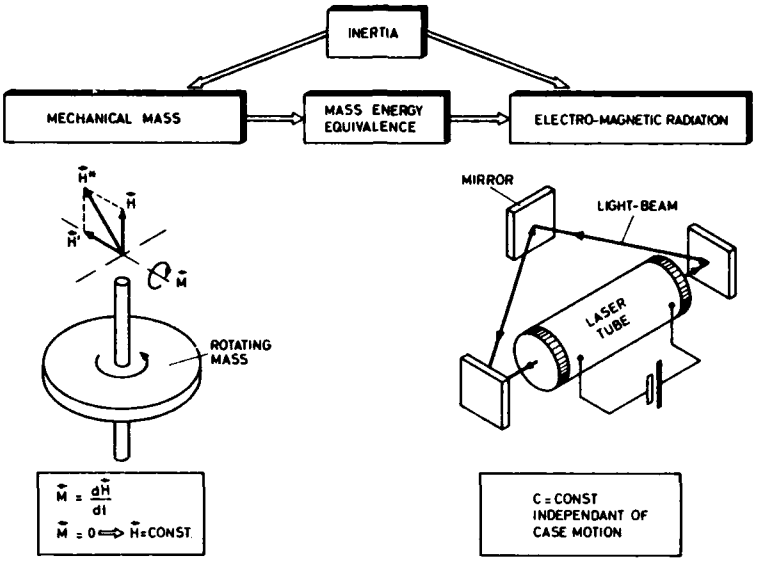
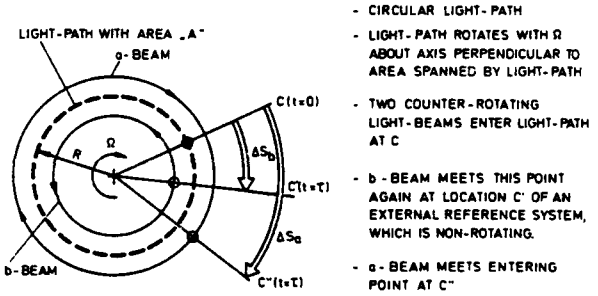


Fig. 2-2: PHYSICAL REFERENCE QUANTITIES FOR INERTIAL RATE SENSING



- BEAM PROPAGATION TIMES FOR NON-ROTATING OBSERVER

$$\tau_b = \frac{2\pi R - \Delta S_b}{C} = \frac{\Delta S_b}{R\cdot\Omega}$$

$$\tau_a = \frac{2\pi R + \Delta S_a}{C} = \frac{\Delta S_a}{R\cdot\Omega} \Rightarrow \boxed{\tau_b \neq \tau_a}$$

TIME DIFFERENCE

$$\boxed{\Delta\tau_N = \tau_a - \tau_b = \frac{4A\Omega}{C^2-(R\Omega)^2}}$$

- TIME-DIFFERENCE FOR ROTATING OBSERVER

$$\boxed{\Delta\tau_R = \frac{\Delta\tau_N}{\gamma}}$$; TIME DILATION FACTOR γ

- FOR PRACTICAL ROTATION RATES $R\Omega << C$ AND $\gamma \approx 1$ RESULTING IN THE TIME DIFFERENCE INDEPENDANT OF THE OBSERVER'S STATE

$$\boxed{\Delta\tau \approx \frac{4A\cdot\Omega}{C^2}} \Rightarrow \boxed{\Delta L \approx \frac{4A\cdot\Omega}{C}}$$

Fig.2-3: SAGNAC-EFFECT

A [m^2]	Ω [°/h]	ΔL [Å]	Δτ [sec]	ΔZ = ΔL/λ
10^{-2}	10^{-2}	$6 \cdot 10^{-8}$	$2 \cdot 10^{-26}$	10^{-11}
	10^{6}	6	$2 \cdot 10^{-18}$	10^{-3}
1	10^{-2}	$6 \cdot 10^{-6}$	$2 \cdot 10^{-24}$	10^{-9}
	10^{6}	600	$2 \cdot 10^{-16}$	10^{-1}

$\lambda = 0.6\,\mu m$, 10^6 °/h = 278°/sec

FOR A = $1\,m^2$, $\Omega = 278$ °/sec $\Longrightarrow \Delta z = {}^{1}/_{10}$ OF A FRINGE

FOR A = $10^{-2}\,m^2$, $\Omega = 0.01$ °/h $\Longrightarrow \Delta z = {}^{1}/_{100000}$ MILLION OF A FRINGE

Table 2-1: MEASUREMENT RESOLUTION OF SAGNAC-EFFECT

FEASIBILITY - PROBLEM

SAGNAC - EFFECT YIELDS VERY POOR MEASUREMENT - RESOLUTION

REMEDIES TO IMPROVE RESOLUTION

- MULTIPLE TURN SAGNAC-INTERFEROMETER
- PASSIVE-RING-RESONATOR WITH MULTIPLE PASS OF ONE LIGHT BEAM
- ACTIVE RING RESONATOR WITH FREQUENCY-SPLITTING

Fig.2-4: FEASIBILITY PROBLEM OF SAGNAC-EFFECT AND REMEDIES

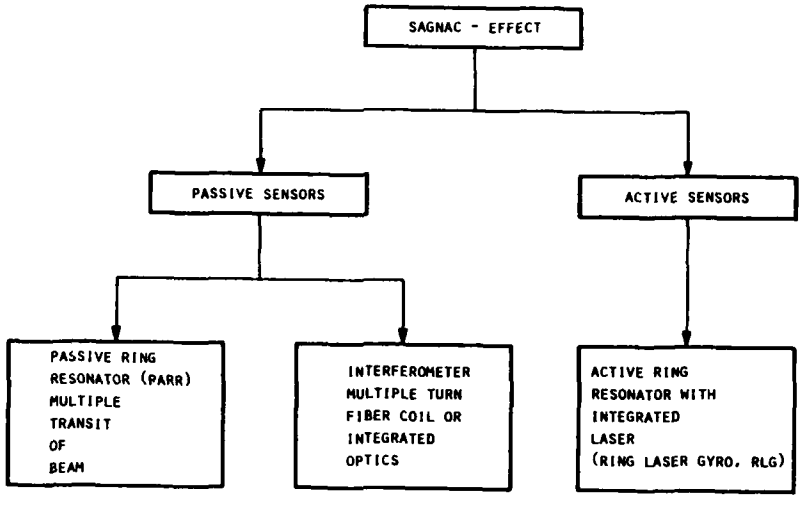


Fig.2-5: ROUGH CLASSIFICATION OF SAGNAC-EFFECT SENSORS

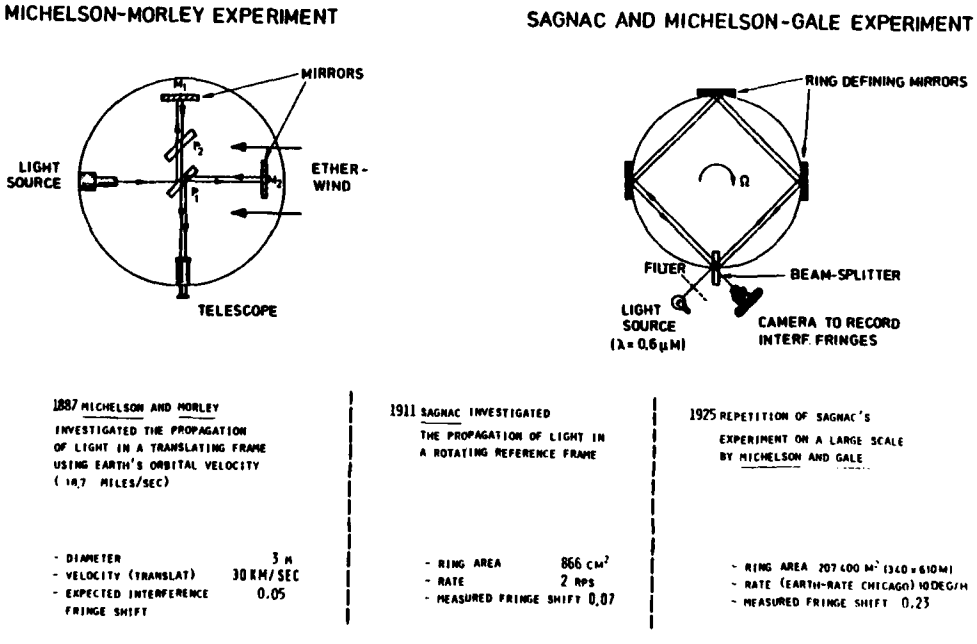


Fig.3-1: HISTORICAL OPTICAL EXPERIMENTS

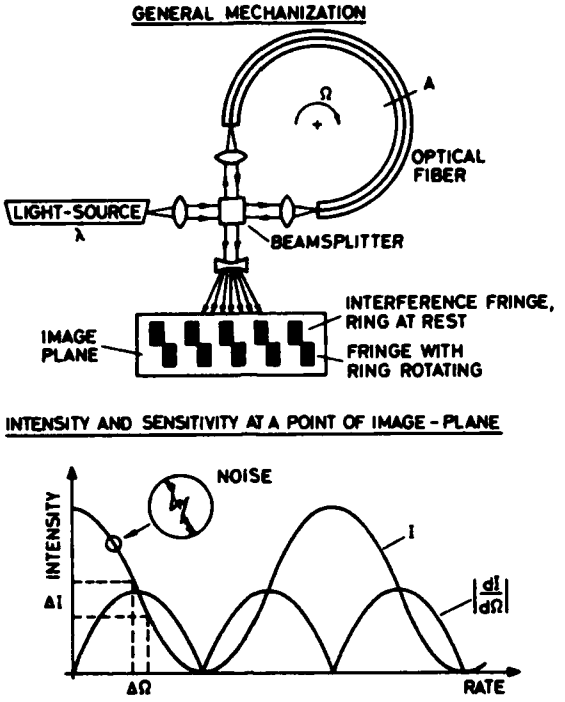


Fig.3-2: SAGNAC-INTERFEROMETER

- SAGNAC PATH-DIFFERENCE OF COUNTER-ROTATING BEAMS

$$\Delta L = \frac{4A}{c} \cdot \Omega$$

- PHASE SHIFT

$$\varphi = 2\pi \cdot \frac{\Delta L}{\lambda}$$

- INTENSITY MODULATION OF FRINGE-PATTERN

$$I = I_0 \left[1 + \cos\left(8\pi \cdot \frac{A}{c\lambda} \cdot \Omega \right) \right]$$

- MAXIMUM SENSITIVITY

$$|\Delta I|_{max} = I_0 \cdot \frac{8\pi A}{c\lambda} \cdot \Delta\Omega$$

- TYPICAL RESULTS

$I_0 = 1\,mW$
$\lambda = 0{,}6\,\mu m$
$A = 0{,}01\,m^2$
$c = 3 \cdot 10^8\,m/sec$

$\Delta\Omega$	$\lvert\Delta I\rvert_{max}$
0,01 °/h	$7 \cdot 10^{-14}$ W
10^6 °/h ≙ 278 °/sec	$7 \cdot 10^{-6}$ W

Fig.3-3: SINGLE TURN SAGNAC-INTERFEROMETER

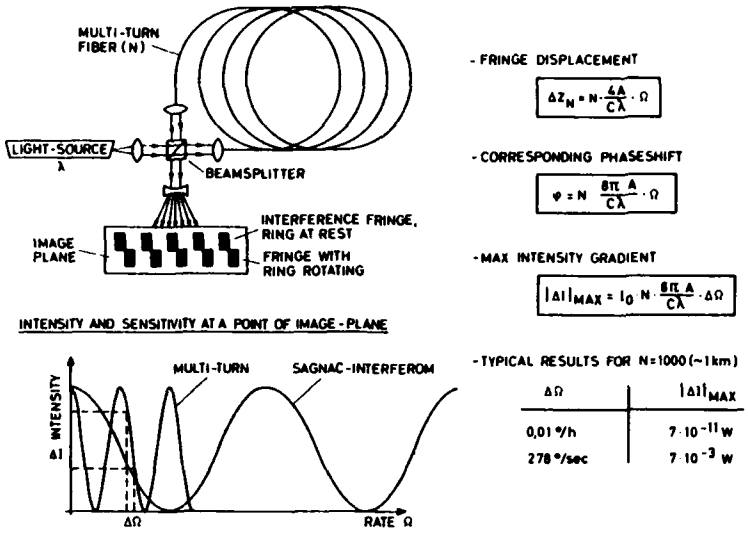


Fig.3-4: MULTI-TURN SAGNAC-INTERFEROMETER (FIBER-OPTIC GYRO)

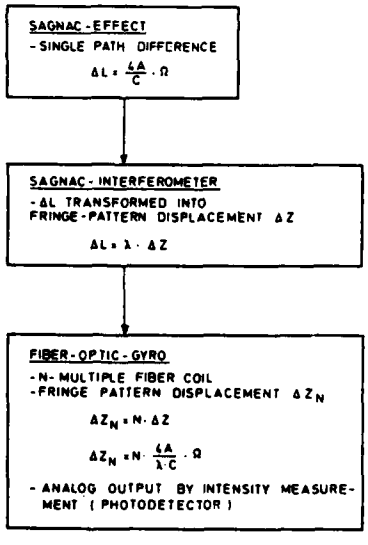


Fig.3-5: SAGNAC-EFFECT TO FIBER-OPTIC GYRO

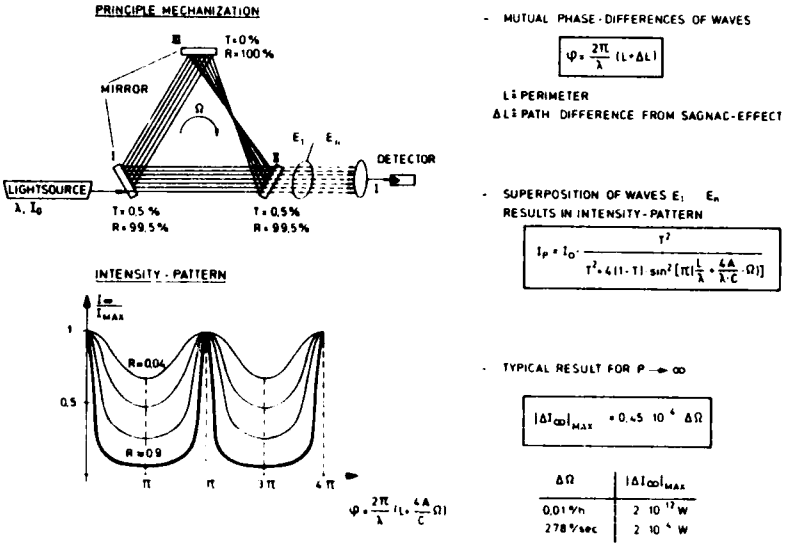


Fig.3-6: PASSIVE RINGRESONATOR (PARR)

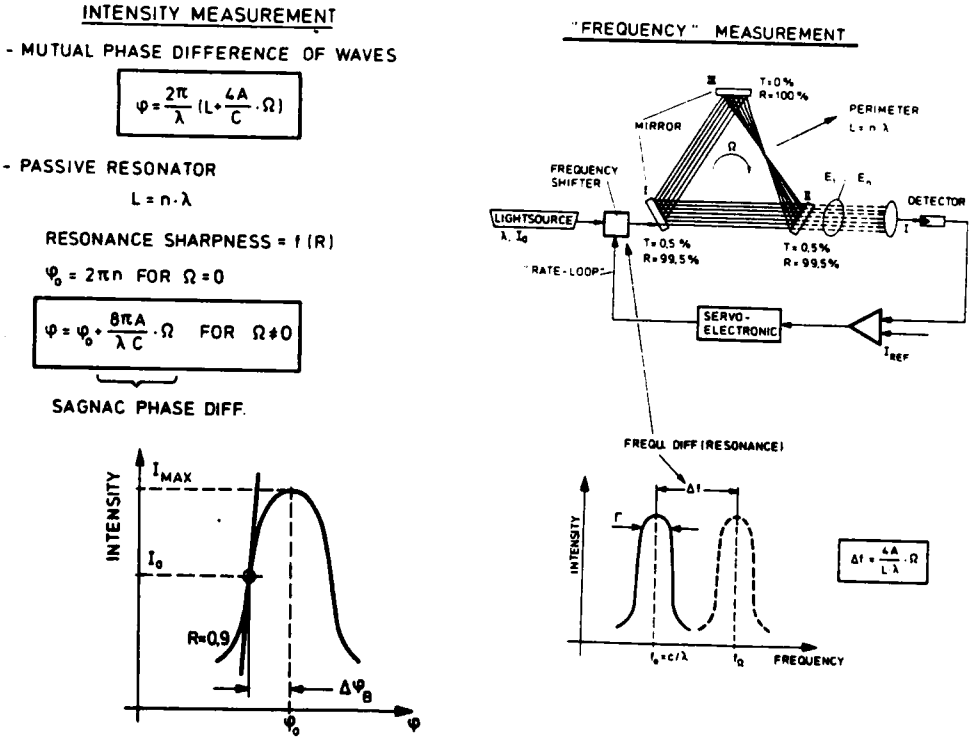


Fig.3-7: "PARR" SIGNAL DETECTION PRINCIPLES

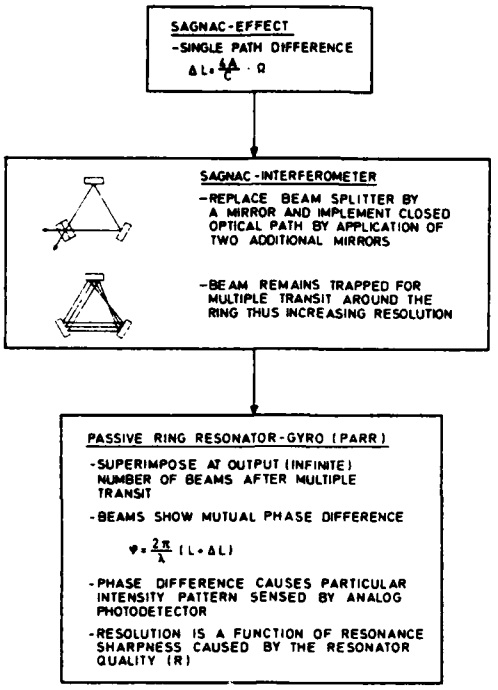


Fig.3-8: SAGNAC-EFFECT TO PASSIVE RING RESONATOR GYRO

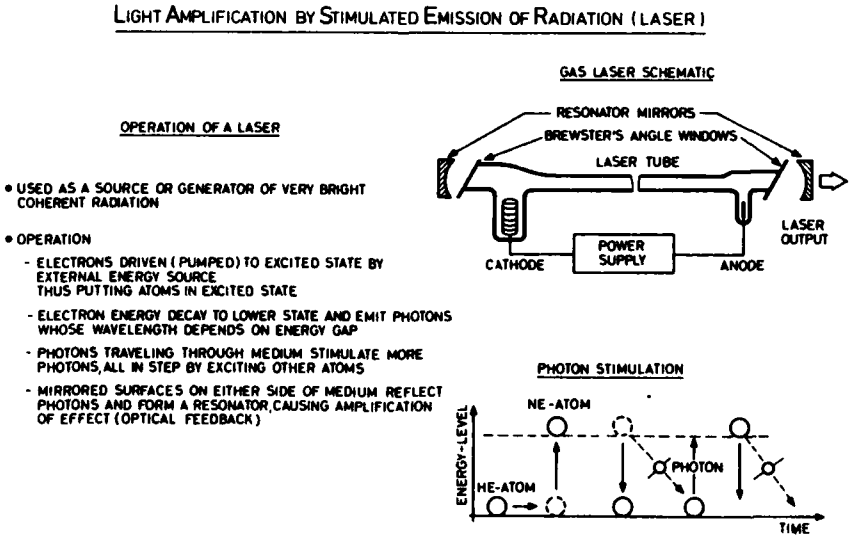


Fig.3-9: WHAT IS A LASER?

OPERATION (CONT'D)

- LASER OSCILLATION CAN OCCUR ONLY AT FREQUENCIES FOR WHICH CAVITY-LENGTH EQUALS AN INTEGRAL NUMBER OF WAVE-LENGTH.

  i.e. THE ROUNDTRIP PHASE SHIFT MUST BE A MULTIPLE OF 2π RAD. SUCH THAT RETURNING WAVE IS IN PHASE WITH INITIATING ENERGY

- RESONATOR IS IN RESONANCE FOR MANY FREQUENCIES. SINCE RESONATOR-LENGTH $L >> \lambda$

- IF FOR PARTICULAR RESONANCE FREQUENCY (MODE) OPTICAL GAIN > LOSSES UNDAMPED OSCILLATION OCCURS AND IS OUTPUT AT MIRRORS (ONE PARTICULAR FREQUENCY).

- LASER FREQUENCY IS A FUNCTION OF CAVITY LENGTH.

OPTICAL FREQUENCY DISTRIBUTION

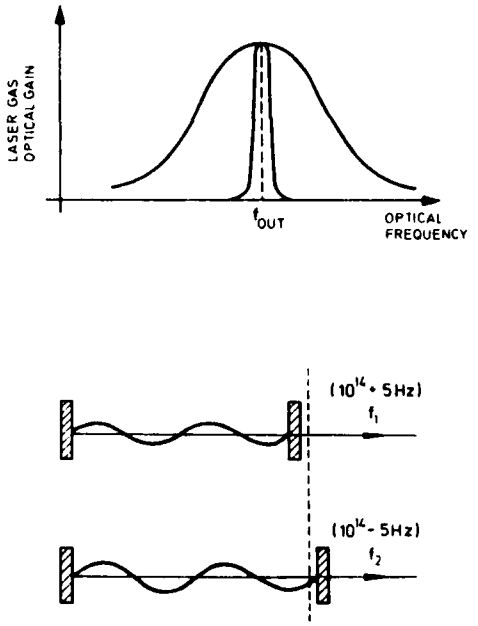
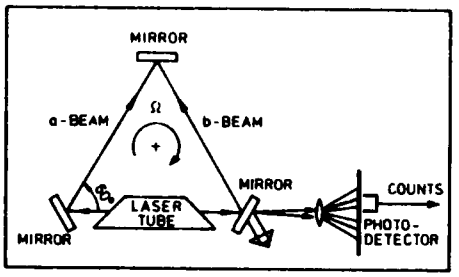
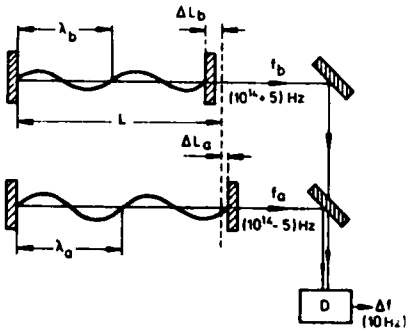


Fig.3-10: WHAT IS A LASER (CONT'D)?

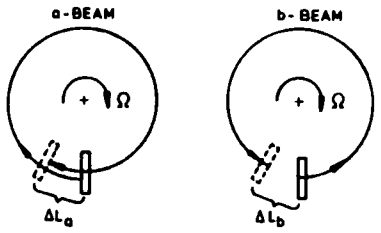
RING-RESONATOR

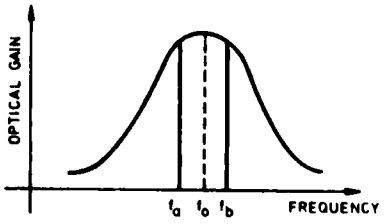


CHANGING RESONATOR LENGTH ΔL CAUSES Δf



RESONATOR LENGTH CHANGES WITH Ω





$$\frac{\Delta f}{f_o} = \frac{\Delta \lambda}{\lambda_o} = \frac{\Delta L}{L} \Rightarrow \Delta f = f_o \cdot \frac{\Delta L}{L}$$

$$\Delta \tau = \frac{4A}{c^2} \cdot \Omega \Rightarrow \Delta L = \frac{4A}{c} \cdot \Omega$$

$$\Delta f = \frac{4A}{L \cdot \lambda_o} \cdot \Omega \quad \text{FRINGE COUNT} \quad N = \frac{4A}{L \cdot \lambda_o} \cdot \Delta\theta$$

Fig.3-11: BASIC FUNCTION OF A LASER-GYRO

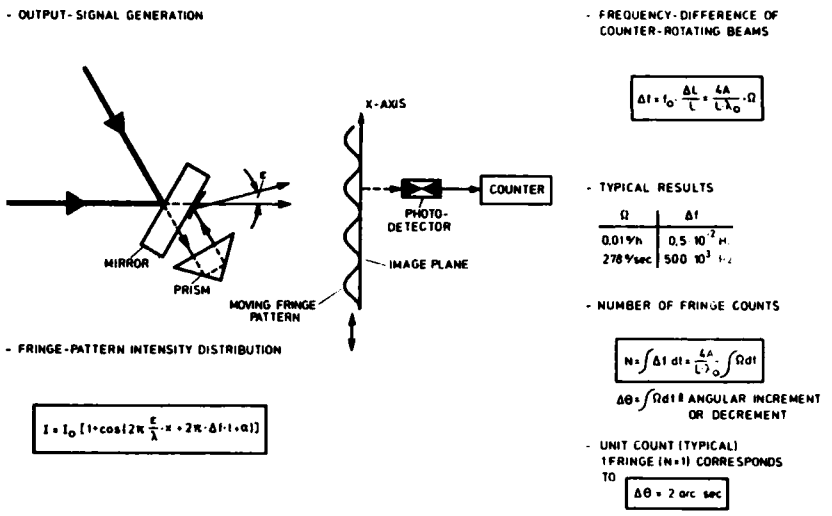


Fig.3-12: RING-LASER-GYRO ⟹ RATE-INTEGRATING-GYRO

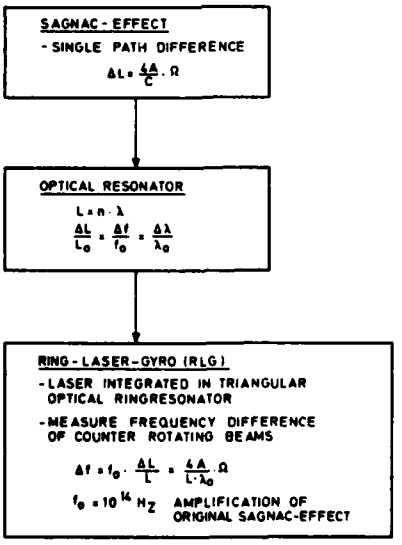


Fig.3-13: SAGNAC-EFFECT TO RING-LASER-GYRO

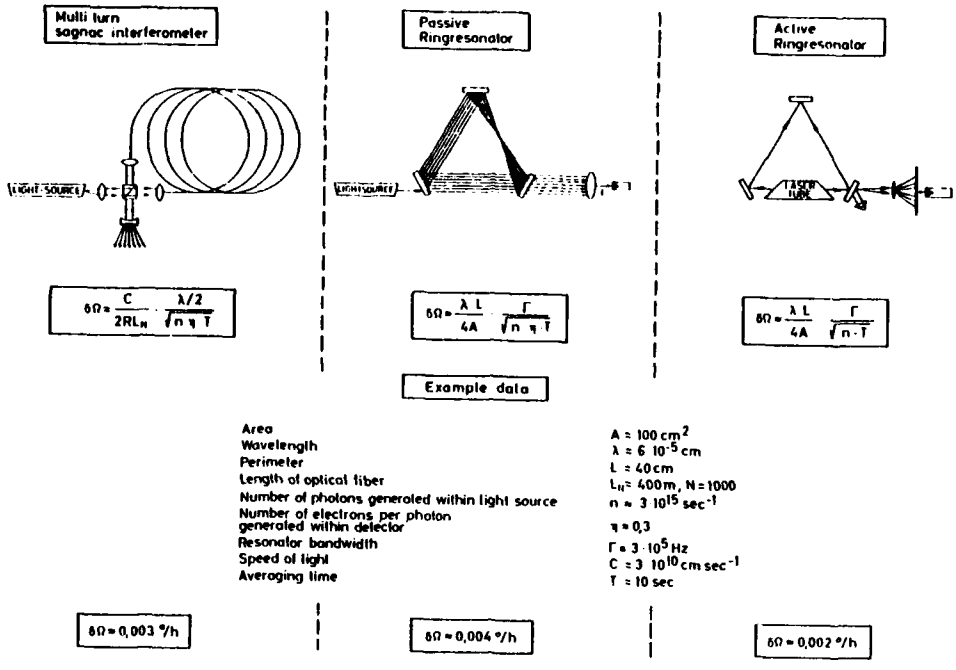


Fig.4-1: FUNDAMENTAL PHYSICAL LIMITATIONS

- OPTICAL RATE SENSOR OFFER CONSIDERABLE POTENTIAL AND WILL PENETRATE INERTIAL-SENSOR MARKET

- ORS'S RATE SENSING REFERENCE TO INERTIAL SPACE IS SPEED OF LIGHT.
  C = CONST INDEPENDANT OF MOVEMENT

- ALL ORSs ARE BASED ON THE SAGNAC EFFECT, WHERE THE PATH DIFFERENCES OF TWO COUNTERROTATING BEAMS FOR $\Omega \neq 0$ RESULT IN AN INTERFERENCE PATTERN SHIFT (INTENSITY MODULATION)

- TO APPLY SAGNAC'S PRINCIPLE AMPLIFICATION OF BASIC EFFECT IS REQUIRED

  - O MULTI-TURN FIBER OPTIC INTERFEROMETER
  - O PASSIVE RING RESONATOR WITH MULTIPLE BEAM TRANSIT
  - C ACTIVE RING RESONATOR WITH FREQUENCY-SPLITTING AND DIGITAL OUTPUT (RATE-INTEGRATING)

Fig.5-1: "ORS"-SUMMARY

INERTIAL REFERENCES BASED ON SINGLE MODE OPTICAL FIBER WAVEGUIDES

by
Willis C. Goss
California Institute of Technology,
The Jet Propulsion Laboratory,
Pasadena, California 91109

SUMMARY

Recent developments in low loss single mode optical fiber waveguides, semiconductor waveguide lasers and waveguide optical and electro-optical components have made a new class of rotational inertial frame references possible. This paper discusses the subclasses of possible fiber waveguide gyro configurations and reviews one approach which is under development at the Jet Propulsion Laboratory. Signal detection and processing, error sources, expected physical and performance characteristics, and present development status are discussed.

1. INTRODUCTION

Current developments in optical waveguide components have made a new class of rotational inertial frame references based on the Sagnac interferometer possible and attractive. Fiber waveguide losses are currently near 4db/km at wavelengths compatible with GaAlAs semiconductor lasers, and less at longer wavelengths, permitting the use of multi-kilometer sensing pathlengths. Many waveguide optical and electro-optical components have demonstrated, including polarizers, isolators, birefringent wave plates, beam splitters, optical switches and phase shifters. Feasibility of fabricating interconnected waveguide components on integrated optical circuit chips has been demonstrated, offering the potential for low cost optical system fabrication through mask-diffusion processes. Fiber optic Sagnac interferometer gyro's offer the potential of meeting a number of inertial reference needs, from very low cost tactical weapon guidance to nearly drift-free aerospace inertial navigation.

2. BASIC CONCEPTS

The basic effect observed in a Sagnac interferometer is that two electromagnetic waves counter-propagating about a closed optical path will experience different travel times if the interferometer is rotating relative to inertial space. The differential travel time can be directly observed as a temporal displacement of the recombined waves, see Figure 1. If the two waves have the same wavelength λ and are traveling with identical polarizations through a polarization-conserving single mode waveguide of length L wrapped on a circular coil form of diameter D, then the observed phase shift due to a rotation component ω about the axis of the coil can be shown to be

$$\theta_r = \frac{2\pi DL\omega}{\lambda c} \tag{1}$$

where c is the vacuum velocity of light. Note the absence of material propagation constants and the direct length dependence which translates into a dependence on the number of turns in the optical sensing coil. A number of fiber gyro concepts are based upon measuring the Sagnac phase shift directly. If the counter-propagating waves have different frequencies, then the phase shift observed at the output will, in general, vary continuously with time. A useful special case is that for which the frequency difference phase shift exactly cancels the rotation induced phase shift,

$$\frac{2\pi nL\Delta f}{c} + \frac{2\pi DL\omega}{\lambda c} = 0 \tag{2}$$

A beat frequency may then be observed which is directly related to the rotation rate

$$\Delta f = \frac{\omega D}{n\bar{\lambda}} \tag{3}$$

where n is the waveguide effective index of refraction and $\bar{\lambda}$ is the mean wavelength. Note that both the Sagnac and the dual frequency phase shifts are linear with waveguide length, so that the observed beat frequency is independent of the length, or the number of turns in the waveguide coil. A similar expression is obtained for a laser gyro, in which the two frequencies are established by the cavity resonances of the counter-travelling waves. However, the two frequency interferometer is not a resonant device; the two frequencies are established external to the sensing path by a control loop which maintains the net phase shift near zero. As a result, the two frequency interferometer is not subject to the laser gyro problem of null rate lockup and the accuracy of the null phase shift control is improved by the use of a long sensing fiber coiled into many turns.

3. DESIGN CONSIDERATIONS

Figure 2 is a schematic of the general form of the single frequency interferometer, showing the relative phase shifts of the various wave components. Note that in the absence of rotation and with a zero value for the non-reciprocal phase bias, the two waves reaching Detector A will arrive in phase and those reaching Detector B will cancel. The function of the non-reciprocal phase bias is to resolve the directional ambiguity and to improve the transfer function slope. Phase bias may be introduced through the non-reciprocal Faraday Effect, or through time-varying electro-optical or stress-optic processes.

At least three design approaches to a single frequency interferometer gyro, differing through their use of the phase bias are possible. As illustrated in Figure 3, the bias phase angle may be dithered about zero to form a discriminant which may be demodulated to form an error signal; the average sum of the dithered phase bias and the Sagnac phase shift can be constrained to zero, thus making the average bias proportional to rotation rate; and the bias may be set to a fixed value of $\pi/2$ and the rotation rate calculated from the differential signal from the two detectors.

Zero mean phase modulated approaches for a strictly harmonic modulation yield a phase demodulated output signal

$$S = kJ_1(2\psi \sin \pi f_o T) \sin\theta_r \tag{4}$$

where J_1 is a Bessel function of the first kind, ψ is the modulation phase angle at peak, f_o is the modulation frequency and T is the light propagation delay through the fiber coil. The constant k is a function of the laser power, transmission of the common path elements, detector responsivity and amplifier gain. The function as written is linear with rotational phase offset near zero. However, significant offsets from zero would be introduced by the presence of small odd harmonic components in the modulation, arising either from the driving waveform or from odd terms in the response of the phase shifter. Self-compensation techniques for scale factor and offset errors do not appear to be possible for this approach, although the approach is attractive through its simplicity.

The second approach is simpler yet, since the output signal is directly obtained from the mean value of the bias. This approach is appealing since the transfer function is determined entirely by the characteristics of the bias element which could be linear. However, currently available waveguide phase shifters have excessive hysteresis and relaxation effects and so cannot be used in this way. Development of a stable phase shifting waveguide element could make the modulated phase nulling approach very attractive, at least for low accuracy, low cost applications.

The third approach, utilizing a fixed $\pi/2$ bias, is not dependent on waveform shape or phase bias element linearity, but instead upon the linearity of the detectors over a limited range of intensities. As will be discussed in a later section, small offsets from the $\pi/2$ condition do little to degrade the accuracy. Assuming an even division ratio at the second divider, a fixed $\pi/2$ phase bias results in signals

$$S_A = S_o(1-\sin\theta_r) \text{ and} \tag{5}$$

$$S_B = S_o'(1+\sin\theta_r), \tag{6}$$

where S_o and S_o' are functions of the laser power, the transmission of the common optical waveguide paths, the detector responsitivities and the amplifier gains. If S_o can be made equal to S_o' then a rotation rate signal independent of these factors may be calculated:

$$\theta_r = \sin^{-1} \frac{(S_B-S_A)}{(S_B+S_A)} \tag{7}$$

One question which remains to be adequately answered for the fixed bias sensors concerns the linearity of the detectors. A departure δ from linearity in the response

$$S = KI^{(1+\delta)} \tag{8}$$

will result in a scale factor error approximately equal to the nonlinearity factor δ. A detector linearity of the order of 10^{-7}, over an intensity range of $\sqrt{2}$ would be required for some applications.

A second question which has been raised regarding the entire class of single frequency analog sensors concerns the size of the A/D converters which are available. Approximately 21 bits are needed to adequately represent the signal for a high accuracy application, yet available converters with adequate speed are currently limited to about 16 bits. The solution appears to lie with data averaging. Data will be available at a rate equal to the inverse of the light travel time through the interferometer, typically 100 kHz. The demand for data updating will be at a much lower rate, for example, 100 Hz. Assuming the signals have symmetrical noise spreads of adequate width to span several converter bits (either intrinsic or supplied) then the signal centroids may be calculated in double precision to an adequate number of bits.

In contrast to a single frequency interferometer sensors which produce analog signals related to rotation rate, dual frequency interferometer sensors are intrinsically digital, generating beats which represent discrete rotational angle increments. Several design approaches are feasible which principally differ in the way in which the counter-circulating frequencies are established. As noted earlier, the requirement is that the net phase shift resulting from rotation and from the difference frequency remain zero. The two general techniques that may be employed are to excite circulating resonant modes and to vary the beat frequency between two non-resonant frequencies to null the signal observed by an analog phase measurement. Both active and passive resonant cavities are possible, at least in principle, and any of the analog phase measurements described in the single frequency interferometer section may be used in conjunction with a variable frequency oscillator to implement the analog phase null sensor.

Figure 4 illustrates the organization of a design approach which combines the direct beat frequency readout of the dual frequency phase nulling interferometer sensor with the multiple turn coil sensitivity of the single frequency phase sensing interferometer.[1] Acousto-optic modulators internal to the interferometer introduce different frequency shifts f_1 and f_2 to the counter-propagating waves; this frequency difference, as before, introduces a phase shift which nulls the rotation-induced phase shift. Prior to mixing and detection, each wave passes through the "other" acousto-optic modulator; both waves then have the same frequency, offset by $f_1 + f_2$ from the initial laser frequency. Frequency dither provides a phase modulation of the detected signal and a phase sensitive demodulation generates an error signal about the null phase condition. This approach appears to be very promising, if the acousto-optic phase shifters can be coupled to the waveguides without disturbing the single mode propagation characteristics upon which the sensor reciprocity depends.

Several design problems which are common to all these approaches are currently receiving a great deal of attention. These problems relate to a failure to maintain adequate reciprocity through the waveguide paths. Fiber waveguides are not available which adequately preserve the input polarization state, and polarizers used with currently available fibers to select a single polarization state must provide a polarization ratio which is not generally available. In addition, because of the finite propagation time through the sensing coil, a number of environmental effects can introduct nonreciprocal offsets and noise. [2,3,4]

4. JPL SINGLE FREQUENCY INTERFEROMETER

A design approach to a single frequency Sagnac interferometer gyro which is being investigated at the Jet Propulsion Laboratory is shown in block diagram form by Figure 5. This approach uses a pulsed GaAs laser, a fixed $\pi/2$ phase bias, dual silicon photovoltaic detectors, a two kilometer fiber coil and a double-pole double-throw optical waveguide switch which modulates the Sagnac phase shift by periodic reversal of the connections to the sensing coil. A 16 bit microprocessor along with an arithmetic chip to perform the trigonometric function calculations, detection circuitry and drivers for the laser, optical switches and phase bias complete the sensor. The laser is operated at a 50 kHz rate with a duty cycle approaching one half. The signal modulation switch is operated at 25 kHz and the detectors are biased and gated off between return signals to minimize response to reflections and backscatter.

Figure 6 illustrates the signal processing. Signals from the two detectors are first differenced and normalized. Data gathered with the signal modulation switch in the crossed and the parallel states are compared and made to agree by computationally changing the relative detector gains. The two data streams are then differenced, an arc-sin calculation made to obtain the rotation rate and an integrated angular position updated. This signal processing sequence is expected to be self-compensating for errors and variations in laser power, beam splitter ratio, phase bias offset, optical component transmission, detector and amplifier gain and offset. High order residual errors are believed to introduce transfer function slope errors only, without introducing null offset (rate) errors.

Because of the periodic relationship, the calculated rotation rate becomes multivalued for rotation-induced phase offsets exceeding $\pi/2$ and has a measurement uncertainty which becomes large in the vicinity of odd multiples of $\pi/2$. However, by generating a second interspersed data flow resulting from the use of a zero phase bias, an equivalent measurement can be made which has a periodicity displaced by $\pi/2$ and the two measurements combined in a single weighted function as shown in Figure 7. The ambiguities beyond $\pi/2$ for the arc-sin function, and beyond π for the arc-cosine function are resolved by a logical element similar to an incremental fringe counter which checks the value of each function to determine if the other is increasing or decreasing.

An apparent difficulty associated with the signal modulation switch is illustrated by Figure 8. Crosstalk through the switch allows eight light waves to reach the detectors instead of the desired two while the circuitry is gated on, apparently complicating the signal response. However, note that two of the waves have passed through the switch crosstalk twice, leading to a negligibly small amplitude. It can also be shown that the four waves which have passed through the switch crosstalk once can be combined in pairs, with the resultant waves having a zero phase increment. The net effect is introduction of a scale factor error term proportional to the crosstalk ratio, but not the introduction of a null offset.

We believe an essential step in the development of a fiber optic gyro is to arrange the entire optical path, including components, in a continuous solid single mode waveguide. The requirement for reciprocity in a useful sensor is so extreme as to preclude the possibility of separately mounted components. Toward this end, JPL is supporting work aimed at integrating all the optical and electro-optical components other than the sensing coil into a single hybrid waveguide package. The first phase, which has been completed, involved fabrication of a two section alternating beta waveguide coupler and a waveguide phase modulator by titanium diffusion into z-cut lithium niobate.[5] The performance of both elements proved generally to be predictable with symmetrical linear response to a zero-mean field. Crosstalk was quite good at -30 db relative to the throughput; however throughput losses were excessive at -11 to -20 db. Figure 9 illustrates signals from both output waveguides in response to a sinusoidal drive voltage for both TE and TM polarizations. Note that TM mode response is much greater than for the TE, corresponding to a larger effective electro-optic coefficient. Polarization control is obviously required.

Following encouraging results from an interim study of the loss mechanisms, second phase work is now being supported to simultaneously fabricate a number of waveguide components in channel waveguide, as illustrated in Figure 10. All the optical components needed for the JPL single frequency interferometer concept will be on the chip except the laser, the detectors, and the sensing coil. Fiber optic waveguides directly bonded to the channel waveguides will access the devices. Interconnection with the fiber optic sensing coil will be made by electric arc fusion, a process which has been adequately established here.[6]

5. SUMMARY

The basic principles involved and examples of the two principle Sagnac interferometer rotation sensor concepts have been described. A JPL sensor concept has been described.

6. REFERENCES

(1). R.F. Cahill and E. Udd, Opt. Lett. 4 93 (1979).

(2). D.M. Shupe, Appl. Opt. 19 654 (1980).

(3). K. Bohm, P. Marten, K. Peterman, E. Weidel, Proc. IOOC, San Francisco, Calif. (1981) Paper TUL7-1.

(4). R.E. Epworth, M.J. Pettit, Proc. IOOC, San Francisco, Calif. (1981) Paper TuF2.

(5). O.G. Ramer, Hughes Research Laboratory, Malibu, Calif. Final Report JPL Contract 955291, May 1980.

(6). M.D. Nelson, H.T. Fearnehaugh, R. Goldstein and W.C. Goss, Proc. Electro-Opt./ Laser 1979 Anaheim Conference. (Industrial & Scientific Conference Management Ind., 1980).

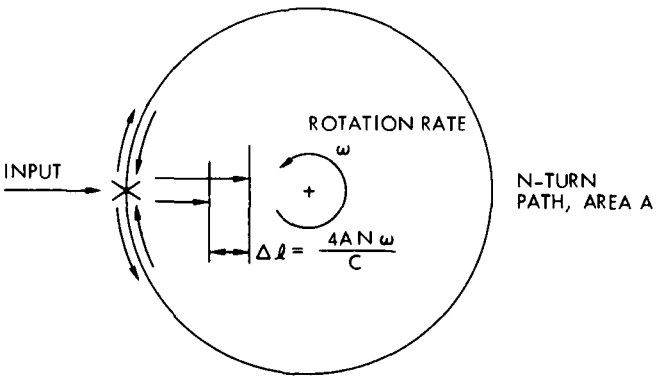


FIGURE 1 SAGNAC INTERFEROMETER DIFFERENTIAL OPTICAL PATHLENGTH

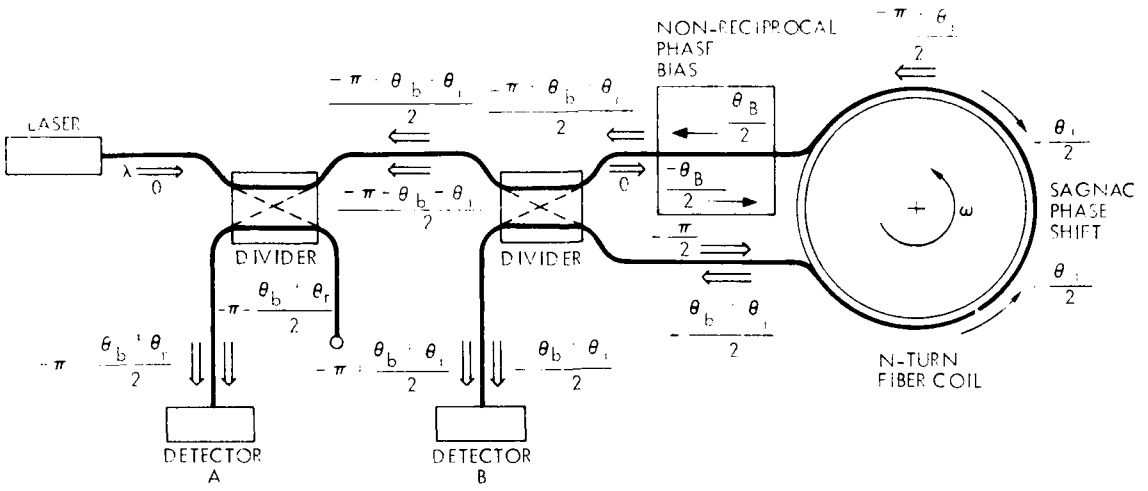


FIGURE 2 SINGLE FREQUENCY SAGNAC INTERFEROMETER PHASE RELATIONSHIPS

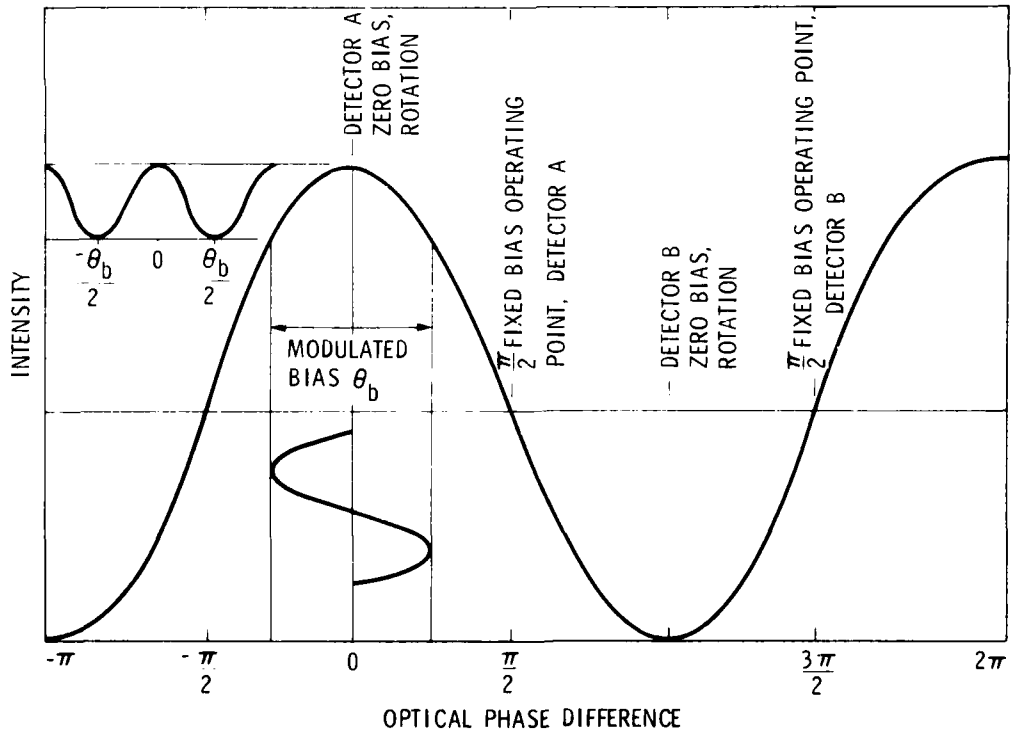
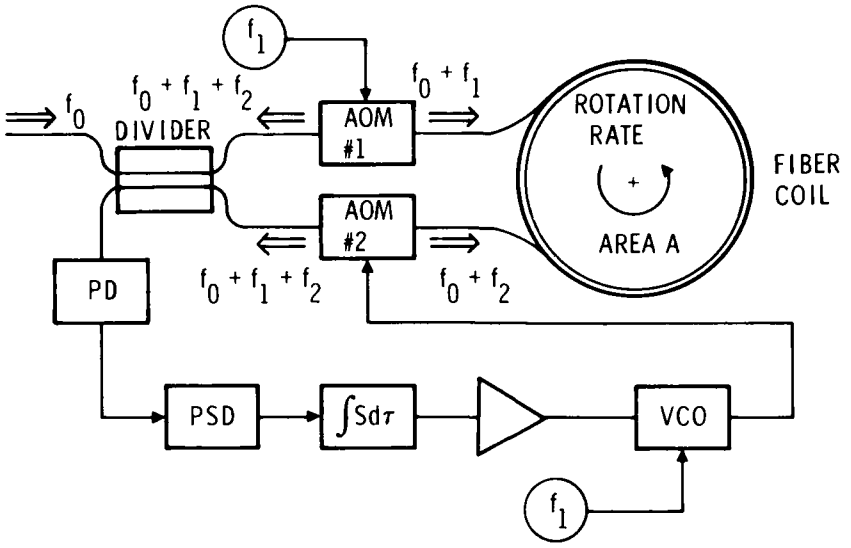


FIGURE 3 INTERFEROMETER OUTPUT INTENSITY VS. OPTICAL PHASE DIFFERENCE



$$\omega = (f_1 - f_2) \quad \frac{\lambda}{d}$$

FIGURE 4 PHASE NULLING DUAL FREQUENCY INTERFEROMETER

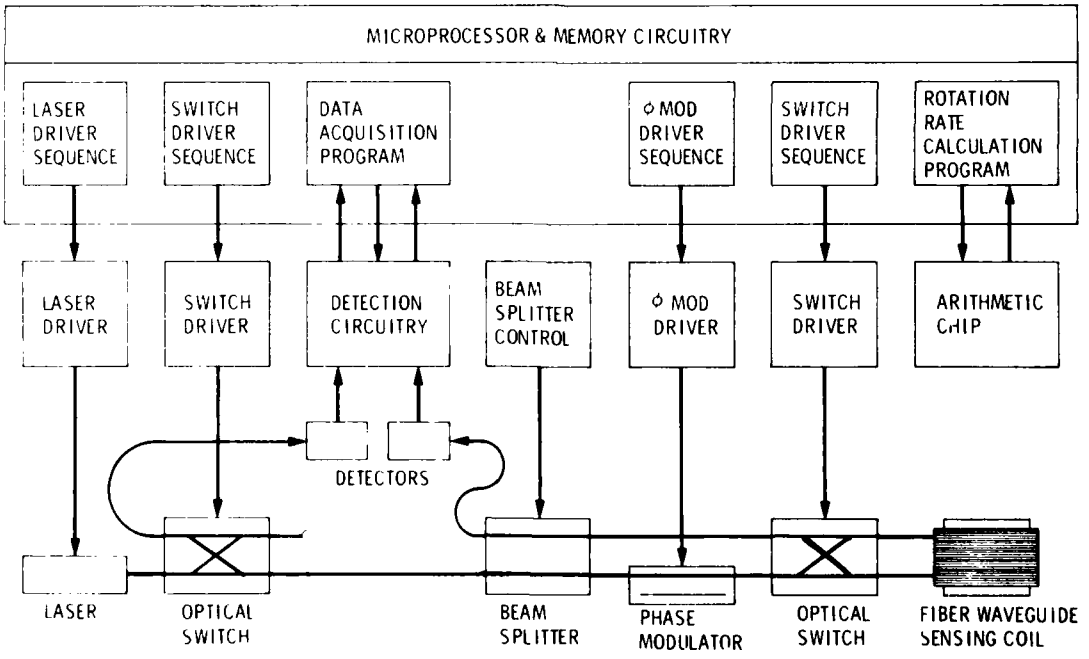


FIGURE 5 JPL SINGLE FREQUENCY INTERFEROMETER

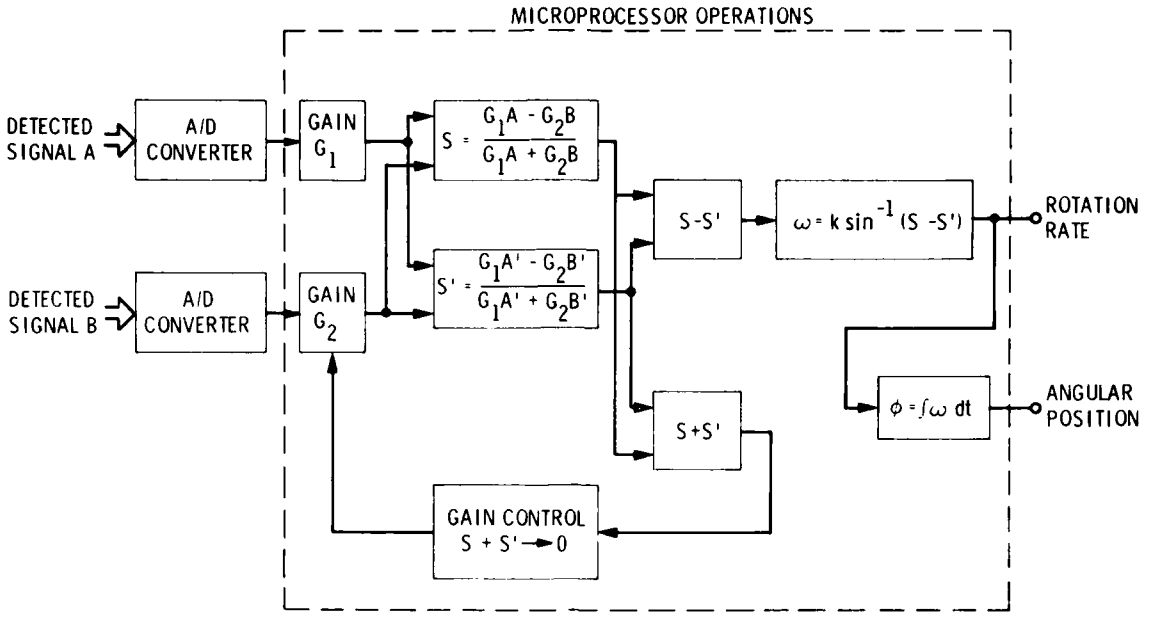


FIGURE 6 INTERFEROMETER LOW RATE SIGNAL PROCESSING

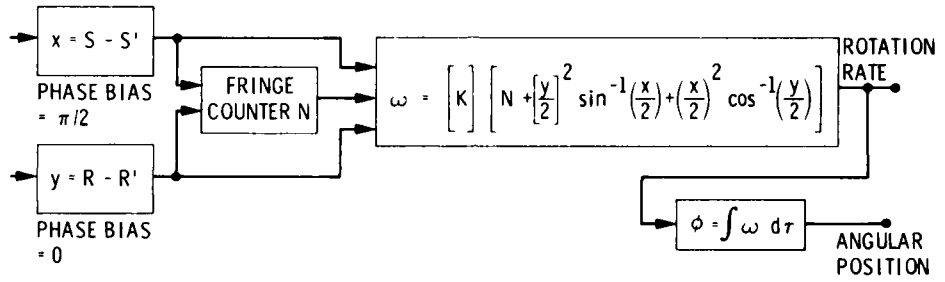


FIGURE 7 INTERFEROMETER HIGH RATE SIGNAL PROCESSING

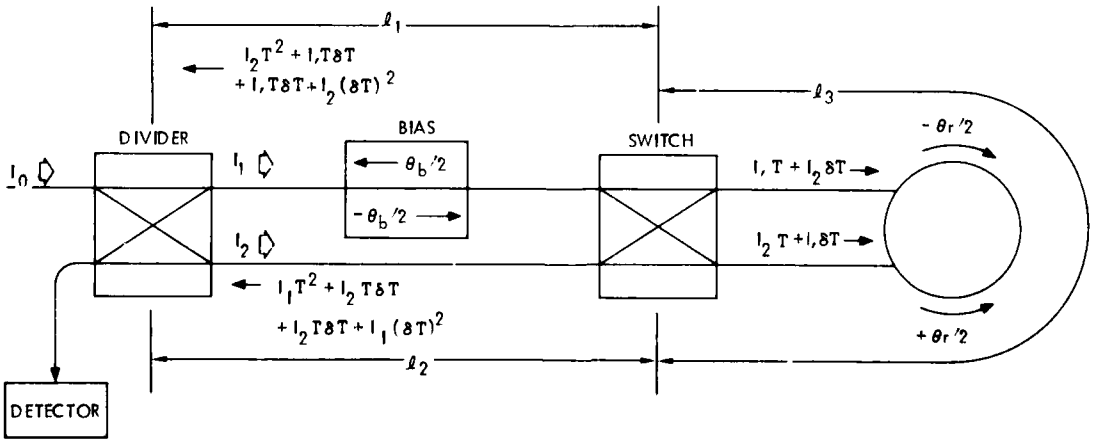


FIGURE 8 OPTICAL SWITCH CROSS TALK

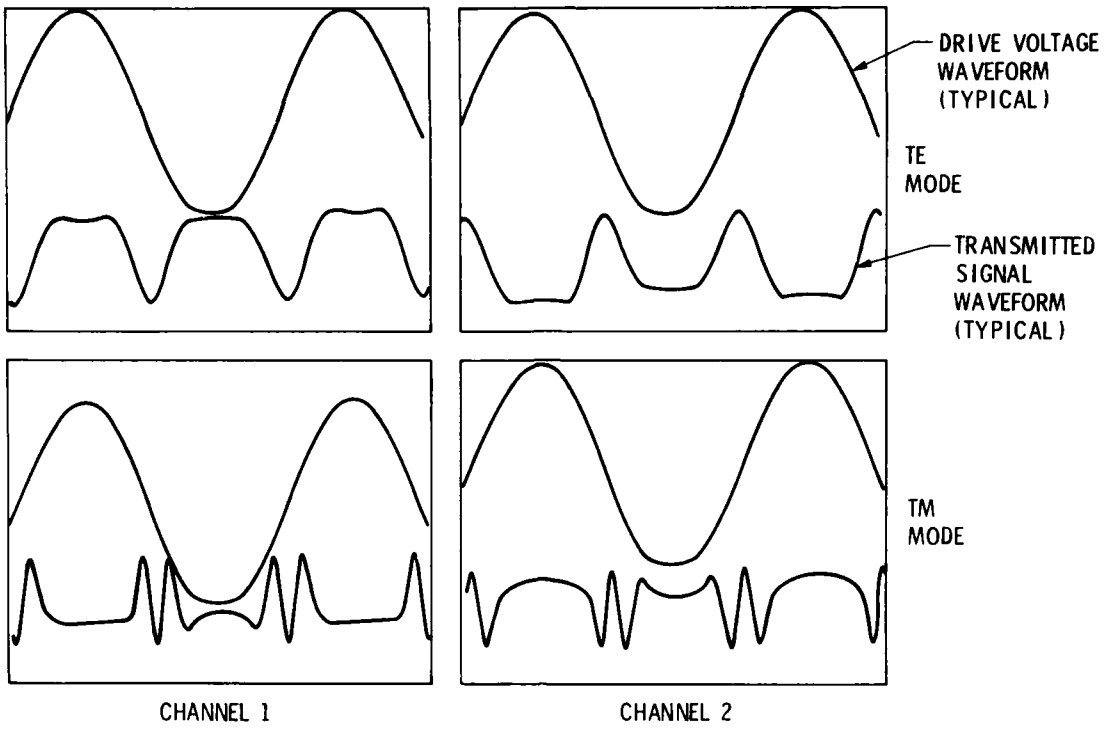


FIGURE 9 WAVEGUIDE COUPLER SINE WAVE RESPONSE

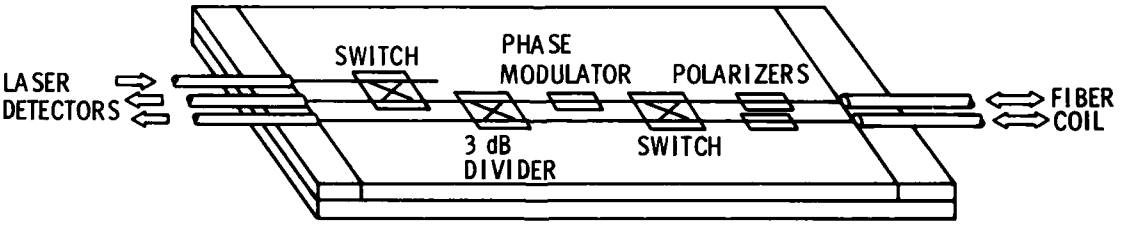


FIGURE 10 INTEGRATED OPTICAL WAVEGUIDE COMPONENTS

LOW COST FLUIDIC SENSORS

H. Douglas Garner
NASA Langley Research Center
Hampton, Virginia 23665

SUMMARY

The gyroscopic inertial sensors used in general-aviation autopilots and in stability-augmentation systems probably contribute more than any other component to the initial cost and to the continuing maintenance costs of these installations. This paper describes several fluidic devices which have been developed to replace the gyroscopes in the conventional "wing-leveler" type autopilots used in small, general-aviation airplanes. These sensors are characterized by simplicity of design, ease of fabrication, and lack of wearing parts. A unique, fluidic, true-airspeed sensor is also described. All these devices are adapted to fabrication by low-cost plastic-molding techniques, and their lack of wearing parts promises long, maintenance-free service lives.

1. INTRODUCTION

Automatic flight-control systems have made an important contribution to the safety and convenience of airplane operation. Under adverse flight conditions, even the simplest stability-augmentation system can make the difference between safe flight and disaster. Autopilots and stability augmenters have long been accepted as standard equipment in commercial and corporate airplanes, but initial cost and continuing maintenance problems have seriously limited their use in the small, personally-owned airplanes that make up a major percentage of the U. S. fleet.

The development of simple, no-moving-parts, fluidic, inertial angular-rate sensors as a substitute for gyroscopes has led to the design of experimental gyroless autopilots of exceptional simplicity and light weight. Other fluidic components and systems under development offer the promise of autopilot designs which are independent of the power supplies of the airplane, and have virtually maintenance-free service lives—comparable to that of the airframe itself.

2. FLUIDICS

Fluidics is an engineering discipline concerned with the flow of fluids through aerodynamically shaped ducts and chambers to perform such functions as sensing, signal amplification and processing, and logic manipulation and mechanical actuation, all of which are traditionally handled by electronic and electromechanical devices.

This field of engineering was first organized as a separate discipline in the early sixties by personnel of the Diamond Ordnance Fuse Laboratories (now the Harry Diamond Laboratories) to encompass a series of fluid-operated logic components and amplifiers which they had invented primarily for use in artillery fuse timers. An outstanding feature of these devices was that they employed no moving parts. It was soon realized that a number of previously existing devices also fitted into this category, and the interest which this newly organized discipline generated in the engineering community soon led to the development of many more device concepts.

In some applications, fluidic devices offered the advantages of superior environmental tolerance, simplified interfacing with the phenomena being sensed or influenced, or improved reliability, resulting from the elimination of moving or wearing mechanical parts. These advantages were frequently offset by the difficulties of providing a suitable source of pressurized fluid, the frequency-response limitations of fluidic components, or the lack of commercial sources of suitable fluidic components.

At this time, much of the original enthusiasm for fluidic devices and systems has waned, and only those concepts offering really unique advantages in particular situations are accorded any consideration.

3. INERTIAL ANGULAR RATE SENSORS

Two fluidic sensors which created early interest in the flight-control area were the vortex rate sensor and the laminar-jet rate sensor. Both of these devices produce output signals proportional to their rate of rotation in inertial space, as does a rate gyroscope, and, since both operate without recourse to moving mechanical parts, they appear to offer advantages in environmental tolerance, extended service life, and simplicity of manufacture over the conventional gyroscope.

The first of these devices to receive attention was the vortex rate sensor. It consists of a flat, cylindrical chamber into the periphery of which a flow of fluid is introduced through a porous wall or a narrow slit called the "coupling element." (See figure 1.)

If the device is not rotating, the flow of fluid proceeds in straight, radial lines directly to the outlet hole at the center called the "sink". If, however, the device is rotated, with respect to inertial space, about its symmetrical axis, the fluid, as it passes through the coupling element, receives a component of momentum tangential to the

cylindrical chamber. As the flow proceeds toward the sink, it forms a vortex of increasing angular velocity until it exits through the sink. The rotational velocity of the fluid at the sink is proportional to, but many times greater than, the rotational velocity of the entire device in inertial space. The rotational velocity of the fluid at the sink is converted to a proportional differential pressure by various pick-off configurations, generally in the form of differential pitot-tube arrangements.

Poor signal-to-noise ratios and poor zero stability were the major problems with these sensors. Work at NASA Langley Research Center eventually led both to designs, operating in the laminar-flow region, that showed greatly improved signal-to-noise ratios and to high-gain, low-noise fluidic signal amplifiers (reference 1) that could bring their outputs up to usable levels. A major effort was made to design these devices so that they could be fabricated by low-cost plastic-molding processes. (See reference 2.) However, poor zero stability remained a problem.

The second of these inertial rate sensors, the laminar-jet rate sensor, did not achieve the status of a practical instrument until somewhat later, probably because its output signal level was several orders of magnitude lower than that of the early vortex rate sensors. Like the vortex rate sensor, its priciple of operation is based upon the conservation of momentum in a fluid stream.

Fluid is passed through a nozzle (figure 2) at such a velocity as to produce a jet, operating in the laminar-flow region, that retains a coherent, columnar form for some distance from the nozzle. As the device is rotated in inertial space, the individual molecules of fluid tend to continue moving in the direction in which they moved when leaving the nozzle (the so-called "Coriolis Effect"); the result is a deflection of the jet with respect to the structure of the rotating instrument. This deflection, which is proportional to the rate of rotation of the device, is sensed at some distance from the nozzle by an appropriate pick-off and is converted into an output signal—either a differential pressure or a differential voltage.

The simpler geometry of the laminar jet rate sensor not only reduces the cost and complexity of manufacture but the reduced number of critical parameters simplifies the task of dealing with noise and zero-stability problems. The low-noise, high-gain fluidic signal amplifiers, which had meanwhile been developed, tended to alleviate the disadvantage of low output-signal levels.

4. APPLICATION TO FLIGHT CONTROL

In the mid-sixties, Honeywell, Inc., under contract to NASA Hugh L. Dryden Flight Research Center, developed an experimental autopilot using a number of fluidic components. (See reference 3.) A vortex rate sensor, sensitive to yaw rate, drove a pneumatic aileron servo by way of a series of fluidic amplifiers in the so-called "wing-leveler" mode. (See reference 4.) A mechanical directional gyroscope with a pneumatic pick-off was used as a heading reference. The pitch axis was damped by a second vortex-rate-sensor loop similar to that used in yaw, and an altitude-hold mode using a diaphragm-air-chamber device with a pneumatic pick-off was added. This autopilot demonstrated the feasibility of airplane stabilization by means of fluidic inertial sensors, but further development was not pursued.

Meanwhile, NASA Langley had developed a low-threshold vortex rate sensor and its associated high-gain, fluidic amplifiers, all of which could be economically fabricated by a lost-wax, plastic-casting process. In order to gain flight experience with these components, they were incorporated into a simple wing-leveler stability-augmentation system.

In contrast to the Honeywell/Dryden autopilot, this flight-control-system concept was directed toward an absolute minimum of complexity and was designed to be powered by the existing vacuum system in the airplane. It was hoped that, if successful, this system might be economically applied to lower priced, personally owned airplanes which could not normally support a conventional autopilot.

The system consisted of a single vortex rate sensor, which was positioned so as to be sensitive to both yawing and rolling rates, a series of fluidic amplifiers to bring the differential pressure signal from the sensor up to a level sufficient to power rolling-diaphragm-type servos, and the servos themselves, attached to the aileron control system of the airplane. (See figure 3.)

This arrangement constituted a minimum wing-leveler configuration, which could maintain the airplane in virtually straight-and-level flight during periods of pilot inattention or disorientation. With the subsequent addition of some form of heading reference, this arrangement could serve as a complete two-axis autopilot.

It was anticipated that, eventually, an efficient, low-drag, ram-air powered venturi vacuum source might be developed to relieve the system from dependence upon the engine-driven vacuum pump (reference 5) and that some type of fluidic/aerodynamic interface might be developed to apply control torques to the airplane without use of moving-part servos.

A number of practical problems arose when the system was exposed to the flight environment. Zero stability of the rate sensor was strongly dependent upon power supply vacuum level and required a sophisticated vacuum regulator. Noise from the engine-driven vacuum pump disturbed the operation of the sensor and amplifiers, thus requiring extensive noise filtration. Mechanical vibration of the plastic tubing connecting the sensor and amplifier units caused similar problems and required that the entire assembly be potted in plastic. Pressure recovery of the fluidic amplifier configurations available at that time was not sufficient to drive the commercial diaphragm servos installed in the airplane, with the result that diaphragm-type power amplifiers were required in the final stage of the amplifier chain. In spite of all these problems, successful flight tests were completed in a typical, single-engine, light airplane. It was evident, however, that a considerable amount of development effort would be needed to reduce the complexity and cost of the system to an acceptable level.

At about this time, an extremely simple laminar jet rate sensor concept appeared which incorporated an integral fluidic-to-electrical signal transducer. (See reference 6.) As shown in figure 4, the deflection of a laminar jet of air is used to detect rotational motion, as in the previous example. (See figure 2.) In this case, however, the deflection of the jet is detected by its differential cooling effect on a pair of self-heated thermistors positioned at either side of its path. These thermistors are maintained at a constant temperature by individual bridge circuits. The difference in the voltages required to maintain these two thermistors at their operating temperatures, while being differentially cooled by the deflected jet, is then a measure of the magnitude of the jet deflection and, consequently, of the rate of rotation of the device. The immediate conversion of jet deflection into a voltage signal by this simple mechanism is a major contribution to the simplicity and ease of fabrication of this instrument. This device was easily adapted to the plastic fabrication techniques then in use and was incorporated into a wing-leveler system in which signal amplification and processing were dealt with by electronic rather than by fluidic amplifiers. (See reference 7.)

Although this approach precluded some of the features envisioned for the all-fluidic wing leveler, the substitution of standard, electronic, integrated circuit components for the special fluidic amplifiers used in the original system resulted in a great reduction in size, cost, and operational problems.

A further advantage of the electronic signal-processing system was the opportunity for adding a simple flux-gate magnetometer as a heading reference. (See reference 8.)

The result was an extremely simple, two-axis autopilot containing no gyroscopes and no moving-part components other than the aileron servos. Successful flight tests of this hybrid system, now referred to as the "electro-fluidic autopilot," were completed in 1973. (See reference 7.)

Commercial manufacturers of avionics hardware were, at that time, reluctant to contemplate any type of flight control system not utilizing conventional gyroscopes, but a descriptive article appearing in Sport Aviation, the journal of the Experimental Aircraft Association (reference 9), created a great deal of interest among the homebuilt airplane enthusiasts. It was apparent to this group that the electro-fluidic-autopilot concept offered the opportunity for the construction of simple, inexpensive wing levelers and autopilots suitable for the many low-cost, individually owned airplanes in operation in this country. The many inquiries received from members of this organization resulted in further dissemination of design information in the form of technical forums and additional magazine articles. (See references 10, 11, and 12.)

At the present time a number of homebuilt versions of the electro-fluidic wing leveler and autopilot are in service in both homebuilt and commercially manufactured light airplanes, and several manufacturers of avionics hardware have fabricated models for study. Figure 5 shows a typical homebuilt electro-fluidic autopilot with the cover removed to show the rate sensor and circuit board. The magnetic-heading reference is shown on the left. This version employs an electrical servo, of the type designed for use in radio controlled model airplanes to drive a control tab on an aileron of the airplane. The entire installation, including the servo, weighs approximately 1 pound.

Early rate sensors encountered problems with zero shift because of changes in temperature and atmospheric pressure. Improvements in design and adjustment procedures have reduced these problems to a minimum (reference 11), although current designs are not hermetically sealed and have a maximum operational altitude of about 10,000 feet.

The concept of an all-fluidic autopilot, containing no moving parts and requiring no electrical power, remains attractive; and recent advances in the design of fluidic rate sensors and amplifiers bring it closer to realization. NASA Langley has continued its investigation of a means for applying control torques to the airframe fluidically without moving parts, and a recent development of The Griswold Company (P. O. Box 305, Old Lyme, CT 06371) offers great promise. A flow of ram air, collected by small scoops at the wing tips, is directed by fluidic amplifiers to slots facing both outboard and inboard at the wing tips. Flow from the outboard slot forces the wing-tip vortex farther outboard, thus increasing the effective span of that wing and increasing its lift. Flow from the inboard-directed slot tends to spoil the airflow over the top of the wing tip and reduce its lift. Thus, rolling torque can be controlled without mechanical control surfaces, and this control arrangement will interface directly with a fluidic autopilot.

At least one American company is producing a version of the laminar-jet rate sensor intended for applications where high shock loading, rapid turn-on and long storage life are important. (See reference 13.) This device is hermetically sealed to avoid problems from changes in atmospheric density and contains its own vibrating-diaphragm fluid pump. Because of the sophisticated nature of its design, however, it can hardly be classified as a "low-cost" sensor.

Figure 6 shows a simple version of the laminar-jet rate sensor, designed for low-cost autopilot applications. (See reference 11.) This instrument has a linear range of about ± 20 degrees per second, but its output is limited to ± 5 degrees per second by the associated signal processing electronics for autopilot applications. It has a time constant of about 0.5 second which is due primarily to the thermistor pick-off. The thermistors used in this example are 0.043 inch (1.09 mm) in diameter. The total power requirement for this sensor, including the pick-off electronics and the motor-driven air pump used to supply the jet nozzle, is about 30 milliamps at 12 volts. The complete, two-axis autopilot shown in figure 5 requires 86 milliamps at 12 volts, exclusive of the servo. Typical model-airplane servos used with this autopilot require a peak current of 500 milliamps, whereas the solenoid valves used with a vacuum-servo version require about 100 milliamps.

The motor-driven air pump shown in figure 6 violates, to some degree, the criterion of no wearing parts. To alleviate this situation, a vibrating-diaphragm air pump has been developed for this device. This air pump is similar to that used in reference 13 but employs a commercial, miniature loud speaker as a driving element. (See figure 7.) Either of these arrangements may be installed in a hermetically sealed housing to avoid sensitivity changes resulting from variation in atmospheric pressure, but, for light-airplane applications, the added bulk and complexity are generally not warranted.

5. FLUIDIC-AERODYNAMIC TURN SENSOR

Another low-cost sensor which is adaptable to autopilot applications depends upon both inertial and aerodynamic principles. One version consists of a pair of aspirating probes (figure 8), one near each wing tip, connected by suitable tubing to a differential pressure trandsducer. The probes are designed to produce a negative pressure equal to the local dynamic pressure. When the airplane is turning, the wing tip at the outside of the turn is moving at a higher velocity than that at the inside; therefore, a pressure differential between the two probes results which is a function of the turning rate. The mass of air within the connecting tubing, being acted upon by centrifugal force, contributes an additional differential pressure of approximately the same magnitude.

By positioning the probes at suitable locations on the lower surfaces of the wings, near the tips, an additional component of differential pressure proportional to rate of roll is superimposed upon the basic yawing signal, thus providing the additional damping required for a practical wing-leveler system. This roll-rate signal is due to the difference in flow velocity beneath the two wing tips that results from the difference in angle of attack between the two when the airplane is rolling. (See reference 14.) Thus, a composite signal, made up of components of yawing and rolling velocities appropriate for the input to a wing-leveler system, can be obtained from a pair of simple pressure probes strategically located with respect to the wing of the airplane.

This general concept is quite old and has been applied, experimentally, to both indicating instruments and autopilot functions (references 14, 15, and 16), although it does not appear to have been exploited commercially. NASA Langley has made brief flight tests to confirm the performance of this concept and has sponsored the development of a successful, experimental, artificial-horizon type of indicating instrument based upon it.

6. FLUIDIC TRUE-AIRSPEED SENSOR

A measure of true airspeed is frequently needed for both manual and computer-based navigational procedures. The traditional approach to this requirement is to measure impact pressure, convert this measurement into "indicated airspeed," and then correct this reading for differences in atmospheric pressure and temperature. This operation requires two pressure measurements, one temperature measurement and some nonlinear computation.

The final fluidic sensor to be described produces a signal, consisting of a pulse rate directly proportional to true airspeed, which is not influenced by variations in atmospheric pressure or temperature. This instrument is based upon a long-known but poorly understood device called the "vortex whistle." (See reference 17.)

A diagram of the fluidic, digital true-airspeed sensor is shown in figure 9. When this sensor is exposed to the airstream, the swirl vanes impart a rotational velocity to the air passing through it. This rotational velocity is accelerated as the air enters the reduced section and becomes unstable as it reaches the enlarged section. A precession or "whipping" motion results, the frequency of which is directly proportional to the rate of volumetric flow through the device.

The resulting sonic tone is detected by a miniature microphone and is processed to yield a signal suitable for a computer or an indicator. A study of this device has been made to define the engineering parameters required for its design (reference 18), and experimental devices are being fabricated for flight testing.

7. COSTS

Since the fluidic sensors described in this paper, with the exception of that in reference 13, are not currently being produced commercially, the term "low cost" can only be inferred from the simple nature of the designs and the fact that they are being consistently fabricated by individuals in home workshops. For a short time in 1978, a small company offered a kit of parts for constructing the laminar-jet rate sensor, essentially as shown in figure 6, for about $32.00. A kit for a complete wing leveler was offered for $70.00.

These figures do not, of course, reflect the certification, overhead, and marketing costs that would occur in full-scale commercial production for the avionics market. They do, however, give some indication of the very low cost of parts and materials required for the manufacture of this type of equipment. Realistic maintenance costs cannot be accessed at this time, since the only items in service have been handcrafted by individual operators, but the very simplicity of the equipment and the lack of critical moving parts would predict few problems in this area.

8. REFERENCES

(1) Hellbaum, R. F., and McDermon, J. N.: Experimental Design Studies and Flow Visualization of Proportional Laminar-Flow Amplifiers. NASA TN D-8433, 1977.

(2) Hellbaum, R. F., Phillips, A. R., and Page, A. D.: Monolithic Precision Casting Technique. Paper 70-Flcs-8 presented at ASME Fluidics Conference, Atlanta, June 1970.

(3) Lock, W. P., and Gee, S. W.: Flight Investigation of a Fluidic Autopilot System. NASA TN D-5298, 1969.

(4) Phillips, W. H., Kuehnel, H. A., and Whitten, J. B.: Flight Investigation of the Effectiveness of an Automatic Aileron Trim Control Device for Personal Airplanes. NACA Rep. 1304, 1957.

(5) Benjamin, M.: Drag Free, Low Cost Vacuum Source from NASA. Sport Aviat., Feb. 1978, pp. 11.

(6) Moore, A. G.: Angular Movement Sensing Device. U.S. Patent 3,500,691, issued Mar. 17, 1970.

(7) Garner, H. D., and Poole, H. E.: Development and Flight Tests of a Gyro-less Wing Leveler and Directional Autopilot. NASA TN D-7460, 1974.

(8) Garner, H. D.: Magnetic Heading Reference. U.S. Patent 3,943,763, issued Mar. 16, 1976.

(9) Garner, D.: The Saga of the Plastic Autopilot. Sport Aviat., Mar. 1974, pp. 45-59.

(10) Garner, D.: Construction Notes on Electro-Fluidic Wing Levelers. Sport Aviat., June 1978, pp. 32-37.

(11) Garner, D.: Fine Tuning the Electro-Fluidic Autopilot. Sport Aviat., Aug. 1980, pp. 16-24

(12) Garner, D.: A Magnetic Heading Reference for the Electro-Fluidic Autopilot. Sport Aviat., Nov. 1981, pp. 19-26, and Dec. 1981, pp. 28-32, 51.

(13) Swarts, R. E., Slabinski, R. J., and Lopiccolo, M. T.: Super Jet ™ Solid State Fluidic Rate Sensor. AGARD-AG-254, Advances in Inertial Navigation Systems and Components, Apr. 1981, pp. 4-1 to 4-16.

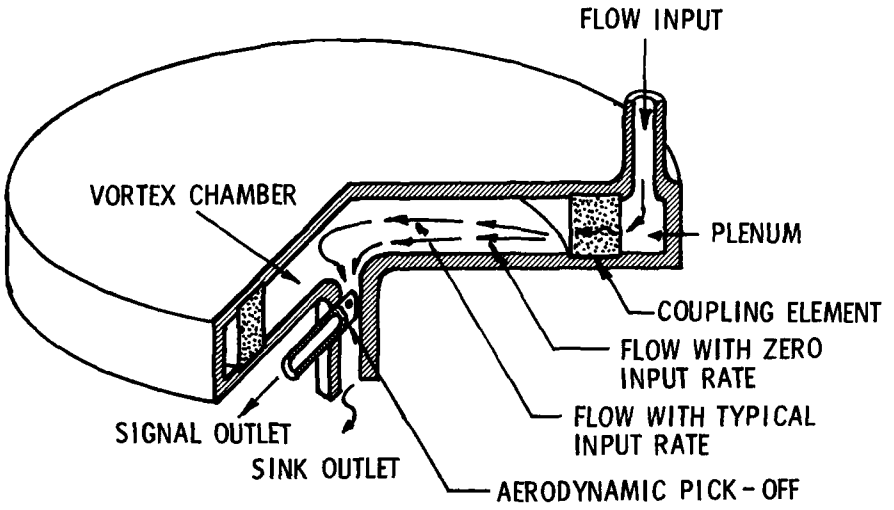
(14) Griswold, R. W. II: Aerodynamic Autopilot. U.S. Patent 3,027,121, issued May 27, 1962.

(15) Darwin, Sir Horance: The Static Head Turn Indicator for Aeroplanes. Aeronautics, Oct. 1919, pp. 410-412.

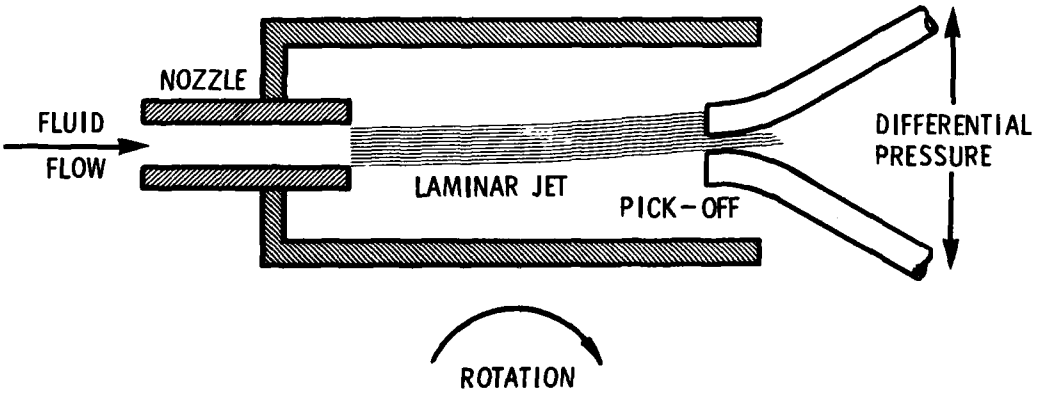
(16) Upson, R. H.: Flight Control by Air Visualization. SAE Journal (Transactions), vol. 30, no. 4, Apr. 1932, pp. 159-170.

(17) Vonnegut, B.: A Vortex Whistle. J. Acoust. Soc. Amer., vol. 26, pp. 18-20.

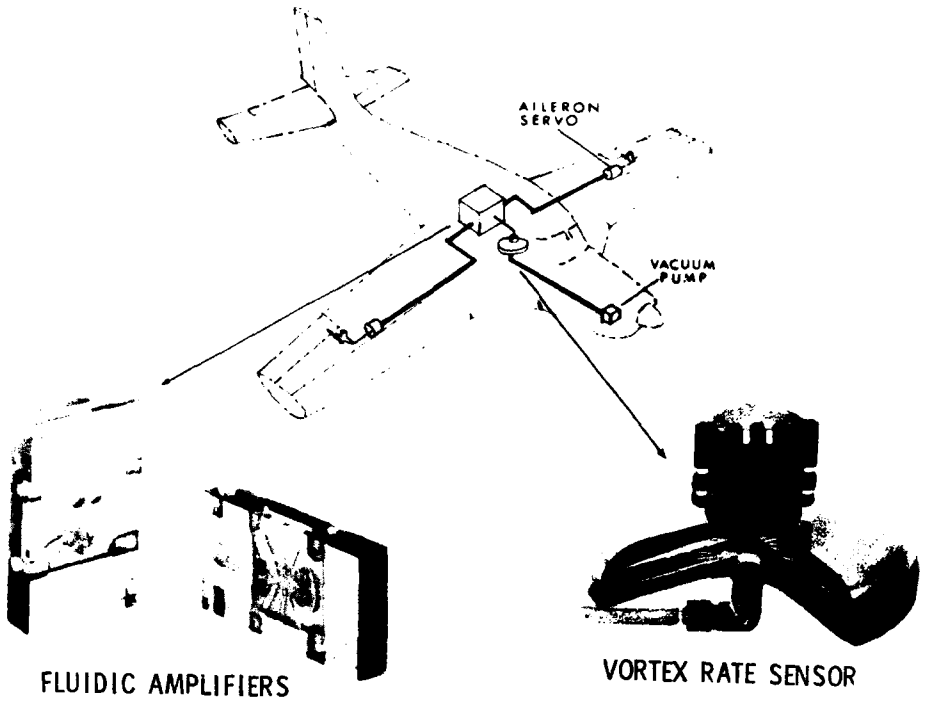
(18) Goglia, G. L., and Shen, J. Y.: Experimental and Analytical Studies of a True Airspeed Sensor. Old Dominion Univ. Res. Foundation, Res. Grant NSG-1177, Final Rep., June 1980.



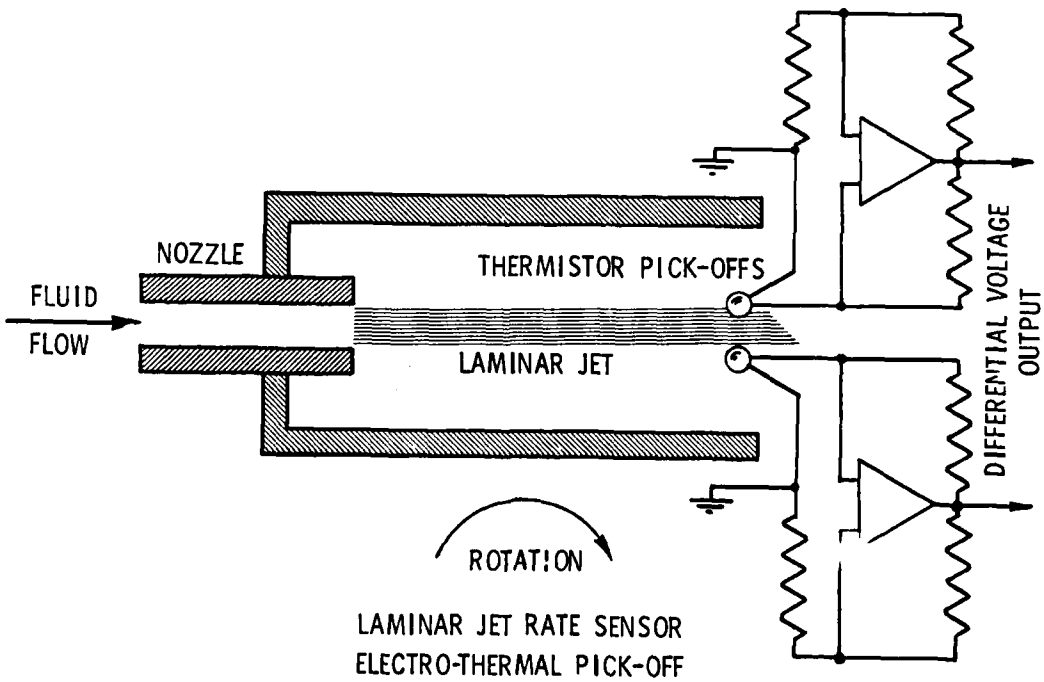
VORTEX RATE SENSOR
FIGURE 1



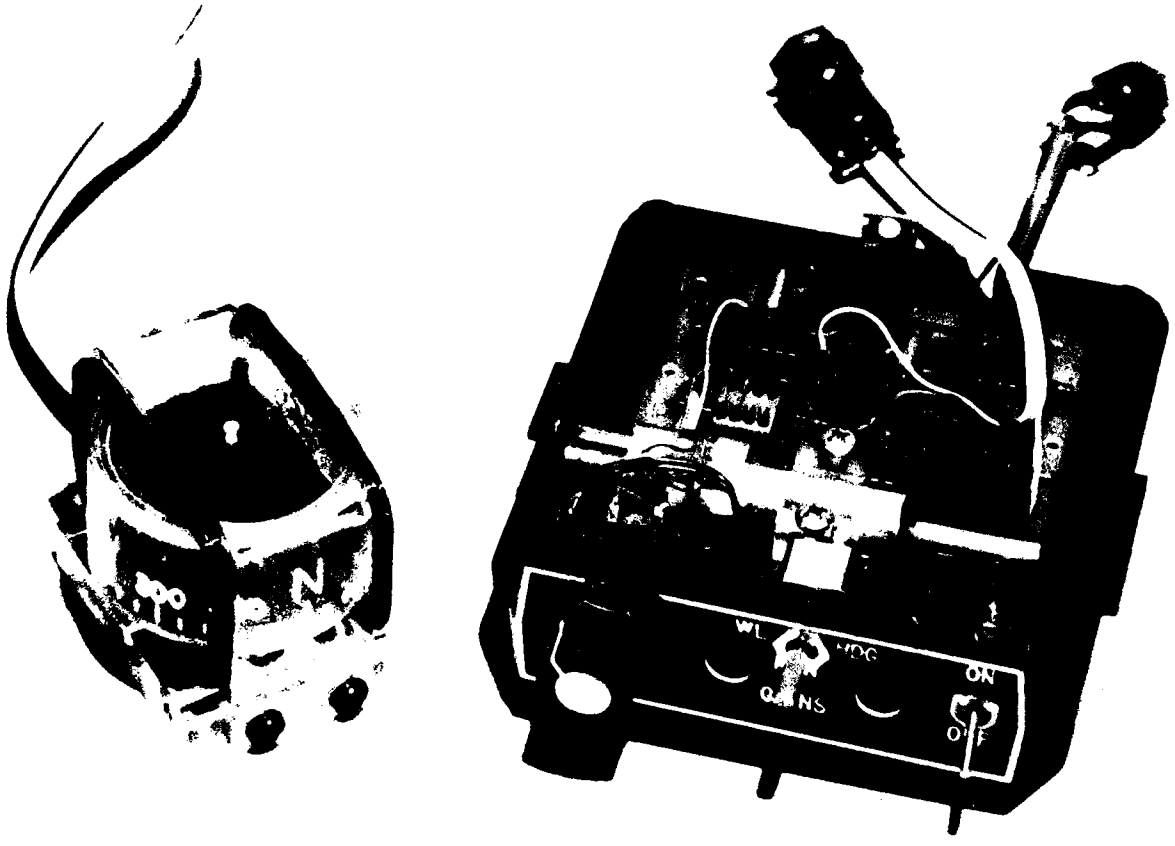
LAMINAR JET RATE SENSOR
DIFFERENTIAL PRESSURE PICKOFF
FIGURE 2



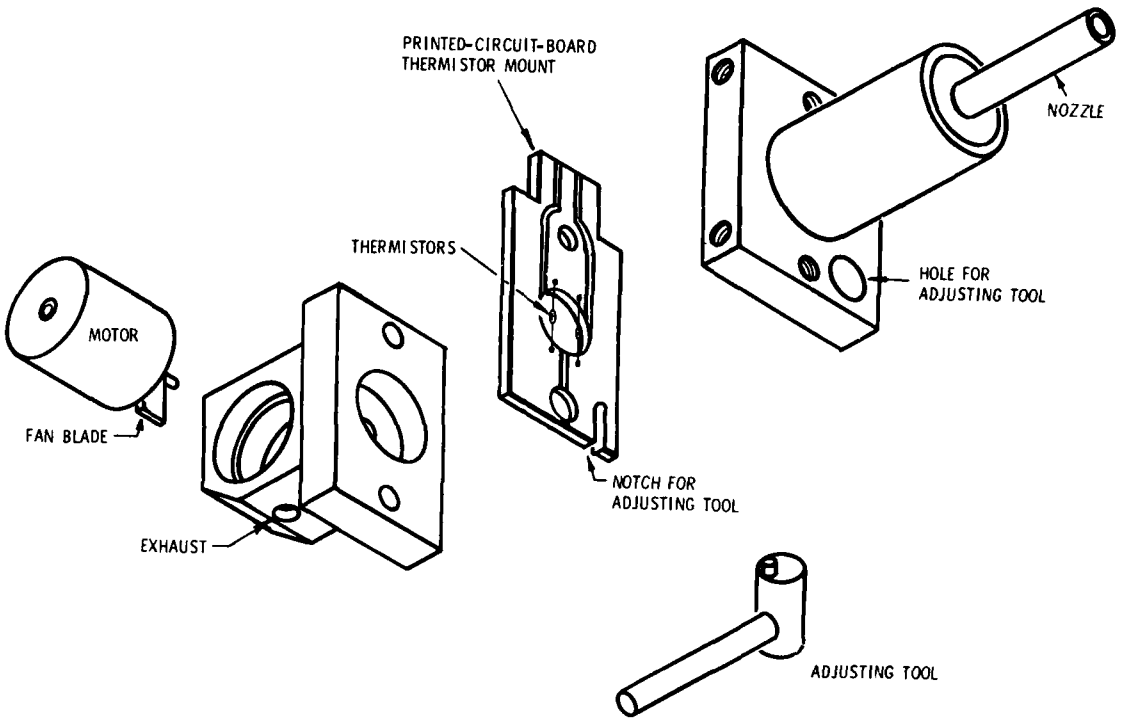
FLUIDIC WING LEVELER
FIGURE 3



LAMINAR JET RATE SENSOR
ELECTRO-THERMAL PICK-OFF
FIGURE 4

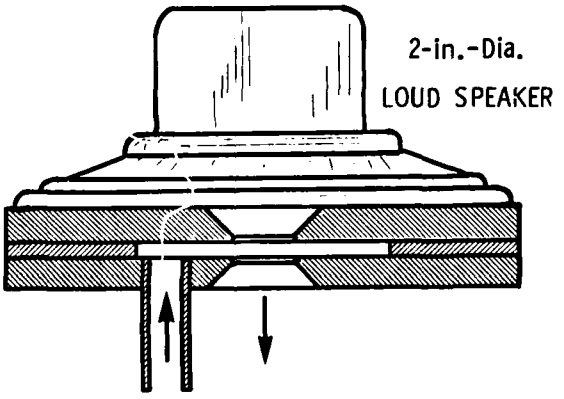


ELECTRO-FLUIDIC AUTOPILOT
FIGURE 5



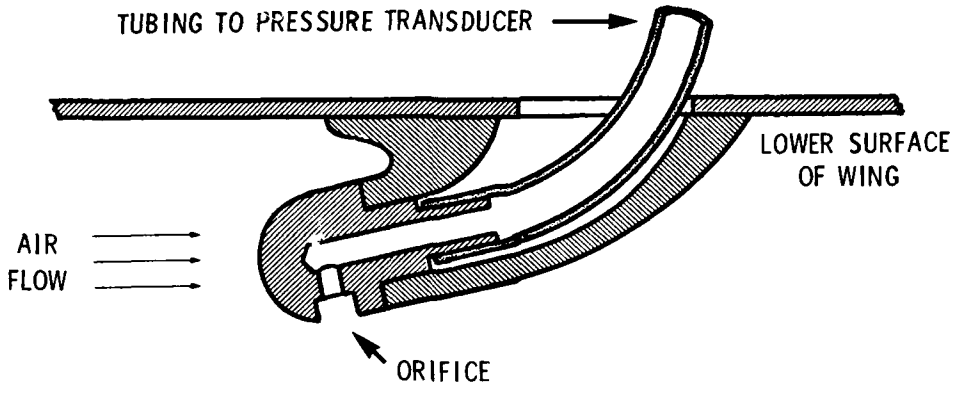
LAMINAR-JET RATE-SENSOR ASSEMBLY

FIGURE 6



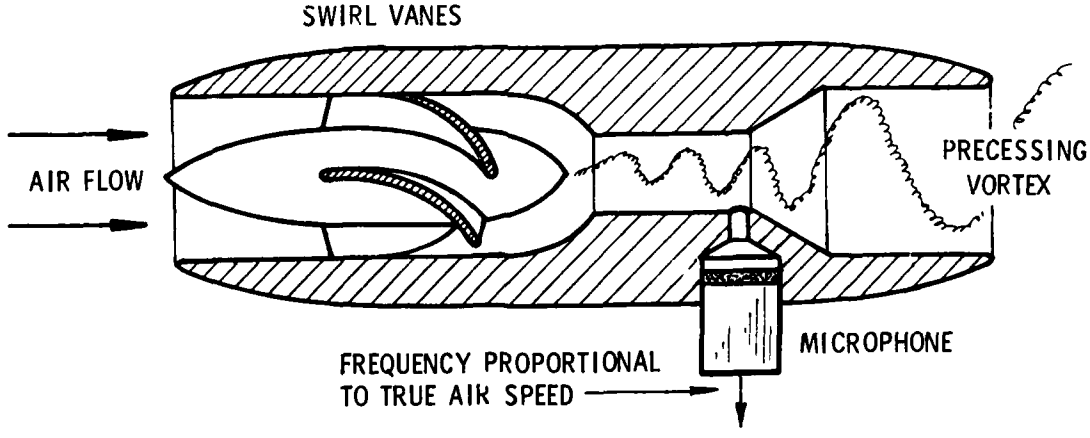
ACOUSTIC AIR PUMP FOR LAMINAR-JET RATE SENSOR

FIGURE 7



AERODYNAMIC YAW / ROLL SENSOR

FIGURE 8

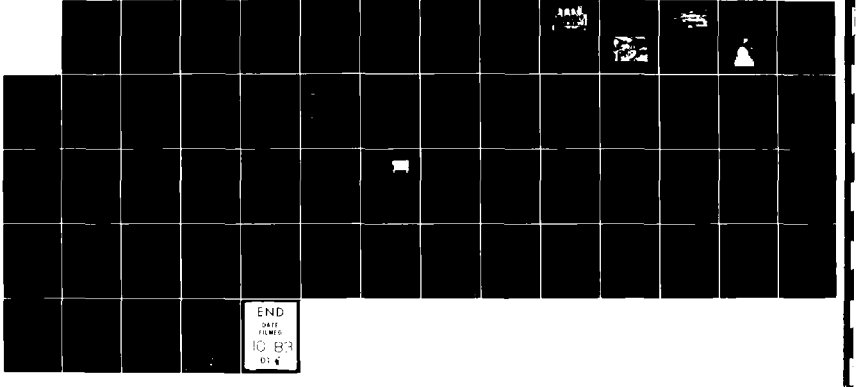


FLUIDIC DIGITAL TRUE-AIRSPEED SENSOR

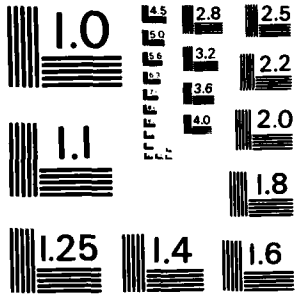
FIGURE 9

AD-A132 910 ADVANCES IN SENSORS AND THEIR INTÉGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS(U) ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY.. 2/2

UNCLASSIFIED J L HOLLINGTON JUN 83 AGARD-AG-272 F/G 1/3 NL



END
DATE FILMED
10 83
DTIC



AIRSPEED AND WIND SHEAR MEASUREMENTS WITH AN AIRBORNE CO_2 CW LASER

by

Alan A. Woodfield
Royal Aircraft Establishment, Bedford MK41 6AE, England

and

J. Michael Vaughan
Royal Signals and Radar Establishment, Malvern, Worcs

SUMMARY

The Laser True Airspeed System (LATAS) developed by the RAE and RSRE and installed on the RAE HS125 research aircraft is described. It has proved exceptionally reliable and rugged. Examples of results are presented including a climb to 43000 ft; flight through a severe thunderstorm wind shear (microburst); pressure error measurements; and signals observed in cloud, heavy rain and from solid objects such as the ground. The paper concludes with some thoughts on other potential applications such as using the sensor for an intelligent autothrottle; for measuring crossflow velocities; for measuring tyre and ground speeds to save tyre wear; and as a combined air data and ground velocity system for helicopters (including a facility to maintain a steady hover).

1. INTRODUCTION

Since about 1970 the RAE have been collaborating with RSRE in the design and testing of CO_2 CW laser doppler systems (sometimes known as laser radars or LIDARS) for wind measurements. The main aim initially was to establish the feasibility of using these systems to make remote measurements of the wind so that very large changes (wind shear) can be identified. The first system tested was ground based (Fig 1) and housed in a lorry with an accompanying caravan to hold the data processing equipment. The system used telescopes with a primary reflector of 30 cm diameter and worked out to ranges of about 1 Km. However, in common with many similar systems, it required frequent attention to maintain the optical system aligned and the laser tuned to the P20 line without any interference from nearby off-axis modes. Despite these problems the system could be kept aligned and stable for a day or two without adjustment, and its effectiveness in detecting wind shear was very good. This led to the development of a compact and reliable system for flight tests in the RAE HS125 aircraft. This Laser True Airspeed System (LATAS) (Refs 1, 2, and 3) has now been flying in the RAE HS125 for 2.5 years and this paper will describe the system and some experiences with it. The system has proved to be both rugged and reliable. The goal of a CO_2 heterodyne system that can be switched on like a light bulb and requires no adjustments has at last been achieved.

The test results have shown that systems of this type could be used for many different tasks in aviation and this paper concludes with the authors thoughts on some of these possible applications.

2. THE LATAS SYSTEM

The various components of the LATAS system are shown in Fig 2. All of these items are of UK manufacture and, although they are for an experimental system, they have been put through many of the vibration and temperature tests required for production equipments. Most of the items were very advanced when specified back in 1978 but are now readily available. The three most important items are the Optics Head, Surface Acoustic Wave (SAW) Spectrum Analyser and the Video Integrator.

The Optics Head was designed and built at RSRE. Its general layout is shown in Fig 3. The main beam from the 3 watt Ferranti Wave Guide Laser is transmitted through a Germanium plate set at the Brewster angle and then focussed at a given distance using a remotely adjustable expanding lens. The returning signal is received through the same lenses and after a change of polarisation it is reflected off the Germanium plate. This signal is then mixed with a small proportion of the main beam and focussed onto the Mullard CMT detector. The detector has a frequency response better than 100 MHz and is cooled to -196° K using high pressure pure air or 'white spot' nitrogen.

The coaxial monostatic system, with the same optics for transmitter and receiver, is particularly convenient and also tends to minimise effects of beam distortion when propagating through an inhomogeneous atmosphere. Detailed studies have been made of the optimum choice of system parameters, and also of theoretical and experimental signal to noise ratio with calibrated standard targets (Ref 4). These show that within experimental error the equipment performs at the ideal quantum limited level with no unaccountable losses. The beam waist, from which the strongest return signals are received can be set at distances up to nearly 300 m ahead of the output lens, which is only 15 cm in diameter. The dc current level from the detector is directly related to the power in the local oscillator beam and provides a convenient measure of the main laser power. The dc current is also used in an electronic locking loop to maintain the laser automatically on the P(20) transition by inserting a very narrow band optical filter, which transmits the P(20) line, in the local oscillator path.

The Optics Head (Fig 4), is shown here installed in the nose of the HS125, where the infra-red beam is transmitted through a 20 cm diameter Germanium window (Fig 5) with a special diamond-like surface coating to withstand abrasion and insect and raindrop impact. The Optics Head also includes the HT power supplies for the laser and the complete unit is only 68 x 32 x 29 cm. Cooling is from a simple radiator and fan assembly and is used, together with a 100 W heater mat, to maintain the laser between about 25° and 45° C. This has proved adequate for ambient temperatures from -60° to 35° C.

The output from the detector travels about 12 m back to the MESL (Racal-Decca) SAW Spectrum Analyser in the cabin of the HS 125. A sample of 25 microseconds is taken by the SAW and transformed within a further 35 microseconds into the frequency plane and transmitted to the Video Integrator. The SAW has a total bandwidth of 25 MHz and with four switchable frequency offsets can cover the frequency range up to 62.5 MHz. 1 MHZ is equivalent to just over 10 kt so this gives a true airspeed range of up to 635 kt, with an ample 50% overlap between each of the four bands.

In the Video Integrator, which was made by Cambridge Consultants Ltd (CCL), the transformed signal from the SAW is converted into a frequency spectrum with 834 channels that are each 30 kHz wide. To improve the signal to noise ratio for low signals the Integrator will then accumulate a succession of SAW samples. The number of integrations may be set between 1 and 16000 on the airborne system. With 100 integrations for a single sample the overall sample rate can be about 160 per second.

In the aircraft system, the spectrum from the Integrator is then processed in another CCL unit to measure the main features of the velocity signature. This summary of the total spectrum and many other items describing the configuration of the laser system are then transferred to displays and the recording system. The form of the signal obtained from the system at various stages is shown in Fig 6. The strongest signal comes from the region where the illumination is greatest and thus the velocity of the peak of the frequency spectrum is usually that of the air at the beam waist. The maximum and minimum frequencies of the spectrum correspond to the highest and lowest velocities within the range where signals are detectable. In the LATAS system these are measured by the spectrum width and average ('centroid') velocity at an adjustable level below the peak. All these measurements together with the height of the spectrum are made in real time on the aircraft. These numbers and the complete spectrum are displayed in the cabin and recorded on the Plessey PV 1513 digital recording system for further analysis. Fig 7 shows a typical velocity spectrum recorded in calm air at a height of 13310 ft. The spectrum is clear and unambiguous. This remains true even for small signal levels and makes the identification of the loss of signal very straightforward.

3. FLIGHT TEST RESULTS

Two aspects of the flight test results are described in this section:

a. system reliability, and

b. results from a variety of different measurements.

3.1 RELIABILITY

A principle aim of the RAE/RSRE research has been to demonstrate that the LATAS system is both effective and reliable as a means for measuring air velocities. Normally one would anticipate that a 'one off' experimental system such as this would be less reliable and require more frequent adjustment than an eventual production version. It was therefore a very pleasant surprise to find that:

a. the system has never needed any adjustment except after changing the laser. (The system has currently been operating for 12 months without any adjustment),

b. the 20 cm Germanium window in the nose of the aircraft, Fig 8, shows no sign of any damage to its diamond-like coating after over 2 years of flying, which included flights through soft hail.

After a few months of initial teething troubles the system has only suffered four failures in 2.5 years. Two of these involved the laser and two were failures of power supplies to the signal processing equipment.

One subjective measure of the reliability of the system is that it has been demonstrated in flight on 15 occasions to outside parties and has never had a failure. The authors believe that a reliable production version of the LATAS system can be made now using current technology and available UK equipment.

3.2 LATAS MEASUREMENTS

Four different types of measurements have been chosen to illustrate the capability of the system. These are:

a. a climb to 43000 ft to show the sensitivity in very clear visibility conditions,

b. a severe thunderstorm wind shear measured during the Joint Airport Weather Studies (JAWS) at Denver, USA in July 1982,

c. aircraft pressure error measurements,

d. return signals in clouds, rain and from the ground.

3.2.1 CLIMBS

If a laser is to be used for measuring essential flight data then it must always produce a usable signal and any failures must be easily identified. Part of the research programme at the RAE has been to find the atmospheric conditions under which the laser is unable to provide a signal and to measure the backscatter coefficient of the atmospheric particles. The signal received at the detector comes from scattering off particles (aerosols), which can be as small as a few tenths of a micron and still produce adequate reflections, but if no particles are present then no signal is returned. Visibility at normal optical frequencies is not always a good guide to either the presence of aerosols or the ability of the infra-red beam from the laser to penetrate to the required distance is, say, fog, rain or smoke. Laser signals have been recorded during climbs to the ceiling of the HS125 (43000 ft) on many flights and Fig 9 shows an example taken on a climb out of Denver in the USA on a very clear Summer morning. The visibility at the surface was over 80 miles and there were no clouds. Despite the clear conditions the laser produced a strong signal near the ground and was only briefly lost during the climb. The strong return at 33000 ft later became a visible layer of cirrus cloud. It is noticeable that the signal increases towards the top of the many small temperature inversion layers and then fails suddenly as the aircraft passes up through the top of the layer. Some of the clearest conditions occur just above the strong surface inversion layer which is fairly common in the UK during the Winter. As height increases further then the signal usually returns. Comparison of laser signals with meteorological data from Radiosonde ascents shows that the laser signal increases significantly when the relative humidity increases.

During the JAWS programme a flight was made together with a Meteorological Research Aircraft from the University of Wyoming, which is equipped with particle measuring equipment. In a climb to 23000 ft the number of particles greater than 0.5 micron fell to zero although the LATAS was still producing a low but clearly detectable signal.

The measurements using LATAS give unique detailed information on the variation of backscatter in different atmospheric conditions. Such information will assist in the initial specification and interpretation of results from the proposed International 'Windsat' meteorological satelite, which plans to use infra-red lasers to measure winds at different levels in the atmosphere.

To summarise, the LATAS system is very sensitive and has always had strong signals within a few thousand feet of the ground. At greater heights there may be regions where no signal is discernible.

3.2.2 THUNDERSTORM WIND SHEAR

Initial interest in the LATAS system was in its potential to meet the need for a sensor to measure large changes of wind before an aircraft encounters them. Action can then be taken to alleviate the dangers of these wind shears. At present wind shear is responsible for the loss of about 2 airliners each year. The most recent was in thunderstorm conditions in July 1982 at New Orleans and killed 153 people (all 145 people in the aircraft and 8 on the ground), and seriously injured 9 people on the ground (Ref 5). The aircraft had just taken off and crashed into a residential area 3/4 nm from the airport. At the time of this tragic accident the RAE HS125 was near Denver, Colorado investigating the nature of the severe wind shears which are associated with thunderstorms. The aircraft with its LATAS system was taking part in the Joint Airport Weather Studies (JAWS) Project which was sponsored by the US National Center for Atmospheric Research (NCAR), the University of Chicago and the US Federal Aviation Agency (FAA). The project has amassed a wide range of data on thunderstorm and other severe weather events from a comprehensive set of ground based sensors, including laser and doppler radars, and aircraft from the National Oceanic and Atmospheric Agency (NOAA), NCAR, the National Aeronautics and Space Administration (NASA), University of Wyoming and the RAE.

Among the many events recorded by the RAE HS125 were several severe downbursts such as that shown in Fig 10. This is the type of wind shear that is believed to have caused the New Orleans, and several earlier accidents. The main features are the very large (nearly 40 kt) and rapid changes of headwind and the strong downdraught of about 1200 ft/min. The downdraught is shown in Fig 10 by the 3 deg increase in the average pitch attitude used to maintain height. If the HS125 had been flying at the much lower speeds used for landing and take-off then this wind shear would have resulted in a large loss of height. Fig 11 shows a model of the flow in a microburst developed from a vortex ring model suggested by Caracena at a Wind Shear Workshop at the University of Tennessee Space Institute, USA in October 1982. This model is able to explain several features that were observed from the HS125 during the JAWS flights, such as dust curtains rising to over 1000 ft around the perimeter of several microbursts. This could not be explained by the more usual vertical jet model which would produce dust blowing radially from the centre of the microburst with very little tendency to rise. The vortex ring model also

produces intense downflows near the ground and, by keeping the energy of the flow contained, it requires less total energy to produce large velocities than the jet model. It also explains the smaller peaks in horizontal velocity observed (Fig 10) at the beginning and end of the microburst at 36 and 54 seconds. This suggests that the HS125 penetrated the microburst just below the upper vortex ring. The aircraft was at a safe height of about 1000 ft above the ground. Fig 12 shows the LATAS system measuring the same wind shear just over 2 sec before it reaches the aircraft, which is equivalent to 250 m distance. The width of the spectrum peak is also interesting as it clearly differentiates between small scale turbulence and significant wind shear. This point is seen more clearly in Fig 13 which shows a sequence of signal spectra from the same time history. By following the sequence it is possible to see each changing wind entering and leaving the tube illuminated by the laser. This tube reaches out to about 700-800 m ahead of the aircraft.

The LATAS sensor has two unique advantages over other airborne sensors:

a. it can measure wind at a known distance ahead of the aircraft, and

b. it can measure the change of wind over distances up to nearly three times its measuring distance.

The range to the beam waist of only 250 m may seem quite short. It corresponds to about 4 sec lead at typical jet transport landing speeds. However, the aim is to provide a system which will enable safe penetration of wind shear. If the wind is measured too far ahead of the aircraft, then any wind shear can change with time or may move sideways relative to the aircraft. Thus there will be only a limited band of distances which will give adequate lead without giving a significant amount of false data. To help identify a suitable distance a laser sensor was simulated during studies of wind shear effects on aircraft. The results showed, Fig 14, that there was a significant advantage in controlling the aircraft using the airspeed measured by a laser at 300 m ahead, but that increasing that distance to 600 m produced no further improvements. On the basis of these tests the maximum range to the beam waist was specified as 300 m for the LATAS system.

The ability of the laser to identify the maximum difference in wind speed over a distance of about 800 m gives it a unique ability to separate wind shear from turbulence before the disturbance reaches the aircraft. This is a capability that autothrottle designers have been seeking for some time. The main problem is that the only speed signal available on all aircraft is airspeed and control of this is essential for large wind variations. But it is not possible to differentiate between large changes and small changes until after they have happened at the sensor. This means that autothrottle systems may on the one hand be sensitive to all airspeed changes, resulting in the discomfort and wear and tear associated with the throttles chasing the smallest changes, or alternatively the system only responds to large changes, making it sluggish. Some improvement can be obtained by using accelerometers to maintain a ground speed until airspeed changes exceed certain limits. However, these systems still require the airspeed of the aircraft itself to change significantly before they respond. The LATAS system can identify a large shear well ahead of the aircraft and even before it reaches the beam waist. This means that throttle activity can be restricted to the conditions where it is really needed without introducing additional lag. A laser system thus provides the signals for an intelligent autothrottle, which will be responsive without unnecesary engine activity.

The wind shear research programme at RAE Bedford is now considering various ways in which the signals from the LATAS (and several more conventional sensors) can be presented to the pilot and automatic control systems and allow safe operations in the presence of severe wind shear.

3.2.3 PRESSURE ERROR MEASUREMENTS

The LATAS system measures airspeed ahead of the aircraft, where it is not affected by the airflow around the fuselage and wings. With the usual air pressure sensors (pitot and static pressure) it is reasonably easy to measure the freestream pitot (stagnation) pressure, but the static pressure measured at the aircraft is usually different from the freestream static pressure. Static pressure (Fig 15) for a given aircraft configuration is a function of freestream speed, pressure, temperature and direction. By very careful choice the static pressure ports can be located where the difference from freestream pressure is small, particularly at normal cruise and landing conditions. However, all height standards for safe separation from the ground and other aircraft depend on static pressure measurements, which are converted into a pressure height. To ensure that all aircraft display the same height scale, they have to be calibrated to measure the pressure error (pe) of each model of an aircraft throughout its full height and speed range.

The present methods of calibration rely on comparison with a reference pressure measured by a calibrated aircraft which uses photography to establish any height differences (or similar flights past ground based tracking towers), or comparison with measurements from a towed static sensor (Trailing Static), which is on a long enough tow to be outside the influence of the aircraft. This latter method requires a major installation on the test aircraft.

The Tower Fly-By at low altitude is satisfactory and convenient, but can only cover a very limited part of the flight envelope. Flights with a calibrated aircraft are costly and difficult (eg the range of speeds of the two aircraft will rarely be the same).

With the LATAS measuring the undisturbed freestream speed and the ability to measure the true stagnation pressure at the aircraft, only a measure of freestream temperature is required to derive the freestream static pressure. Like stagnation pressure the freestream stagnation temperature is also reasonably easy to measure on an aircraft. The freestream static pressure, p_o, is then given, in subsonic flow, by

$$p_o = p_t\left(1 - \frac{V_t^2}{7592}\right)^{3.5}$$

where P_t = freestream stagnation pressure
V_t = freestream true airspeed, kt
T_t = freestream stagnation temperature deg K

Thus from the measured static pressure, P, the fundamental pressure coefficient, C_p at the static ports and its variation with flight conditions of flow angles, Mach and Reynolds numbers can be determined for most of the flight envelopes relevant to different aircraft configurations (eg cruise, landing etc).

C_p is defined as

$$C_p = \frac{p-p_o}{0.7p_oM^2}$$

where M = freestream Mach number

A system such as LATAS could be installed as, say, a temporary replacement for the weather radar and would give complete freedom to establish pe's throughout the full range of configurations and flight envelopes.

Tests with the system on the HS125 are being analysed and some preliminary results at heights from 27000 to 33000 ft are presented in Fig 16. Data at the different heights show a similar pattern with changing incidence angle α but some of the variations with height seem to relate to individual flights. The variation of Reynolds number does not seem to correlate with these height effects. However, the level of scatter is no greater than that seen with other methods. A C_p change of 0.01 is equivalent to only 0.8 mb at the mean test conditions of 0.6 Mach number and 29000 ft height. Further tests are being analysed to establish the individual contributions from Mach and Reynolds numbers. Results will also be compared with those from calibration flights with the United Kingdom Standard (the calibrated F-4 (Phantom) operated by the Aircraft and Armament Experimental Establishment at Boscombe Down) and with Tower Fly-By tests.

The preliminary results with LATAS are very promising. One area where the LATAS system could be particularly useful is in determining true stall speeds and establishing the minimum safe approach speeds for new aircraft. It is essential to maintain a safe speed margin on the approach, but it is very disadvantageous to have approach speeds higher than necessary because of the increased landing distance and tyre wear.

3.2.4 RETURN SIGNALS FROM CLOUDS, RAIN AND THE GROUND

Any system, such as LATAS, that probes the atmosphere must cope not only with a lack of back scattering particles, but also with conditions with high concentrations of particles. If they are very concentrated then they will obstruct the return signal and attenuate it as a function of distance. However, the signal from close range will be very strong. Rain and cloud (or fog) would produce these effects. Another important feature is any non-uniformity in the distribution of back scattering particles. Strong returns from the ground or other solid objects are obvious examples.

High particle concentration has two main effects. First the dynamic range of return signals can be very large as the concentration varies from only a single particle in the beam to a dense cloud. Second the attenuation of the beam and the signal with range will mean that the maximum signal will come from a region closer to the aircraft than the nominal clear air range.

With the LATAS system in its present form, the changes in number of integrations and direct signal attenuation provide a dynamic range of 28 db(630/1). At any chosen value of integrations and attenuation the signal can be accepted with a dynamic range of 20 db(100/1). Finally, the signal is treated as an amplitude signal but the return is really a function of power. Thus the total dynamic range available is about 96 db. This has proved adequate for flight through a wide range of conditions.

In a commercial system the input to the SAW could be controlled by an Automatic Gain Control system and it would be straightforward to optimise the number of integrations.

The attenuation caused by high particle concentration affects the range that a laser system can probe. At the infra-red wavelengths (10.6 micron) of the CO_2 laser the

beam attenuation is small with sub-micron particles, such as the main constituents of smoke. This means that the laser beam can see further than the human eye. This has led to considerable interest in the use of CO_2 lasers in battlefield conditions. However, in well developed fogs, cloud and rain the particles are much larger and in these conditions the laser may not see so much further than the human eye.

As the attenuation increases then the signal returned from different distances changes (Fig 17). With very high attenuation the return signal is very strong but its peak is at the transmitting lens. However, the laser can still detect velocity changes out to the visual range limit and it is still measuring air velocity. A typical doppler spectrum recorded in thick cloud (Fig 18) shows the characteristic saw-tooth return caused by the air in the beam increasing in speed as it approaches the aircraft. (In other locations the air slows down and the saw-tooth would be reversed). The laser is still providing a usable airspeed signal with an error of only a few knots and the characteristic shape of the spectrum could give a basis for a temporary warning that the laser performance is reduced. It should also be remembered that visibility must be less than about 250 m to have any significant effect and probably less than 50 m to produce the effects seen in Fig 18.

In heavy rain the visibility is very rarely less than 250 m and the airflow over the window prevents any significant layer of water attenuating the laser beam. Fig 19 shows a spectrum taken in heavy thunderstorm rain in Colorado, USA and shows no detectable effects other than a strong signal level.

Ground returns can be obtained easily with the LATAS beam set nose down. In the example in Fig 20, taken during a landing approach in tailwind conditions, the ground signal is stronger than the air signal. The relative strengths depend on the slant range to the ground. At long ranges the beam illumination on the ground will be very weak and the air signal will be stronger despite the much higher reflectivity of the ground. If the ground surface is a concrete runway, as in Fig 20, then the ground return spectrum is very narrow as there is no velocity change in the few milliseconds used to sample the doppler signal.

4. FUTURE APPLICATIONS

The remarkable reliability and ruggedness of the experimental LATAS system has demonstrated that commercial 'fit and forget' CW Doppler Laser systems are a practical proposition with the present state of the art. As mentioned earlier, the LATAS has never needed any adjustment or special treatment during over 2 years of flight trials (except for the rare occasions when the laser was changed).

Three possible applications for such systems have already been discussed, viz:

a. wind shear detection at distances of several hundred meters ahead of an aircraft to give time to counterbalance the response lags of the aircraft and its engines,

b. Intelligent autothrottle speed sensor using the ability of the laser to differentiate between turbulence and wind shear at the same time as it measures airspeed,

c. Pressure error measurement without any external influence on the airflow and including measurements during manoeuvres such as the approach to the stall.

Three other possible applications also spring to mind and there will doubtless be many others.

First the LATAS system can be used in a conical scan mode to identify the two crossflow air velocities in addition to the axial flow. The frequency response will be lower but still adequate for most applications.

Second the beam scanning could be arranged to scan the ground and the wheels during landing (Fig 21) and the signals could be used to spin the wheels to give zero slip at touchdown and thus considerably reduce tyre wear. The laser system can give very precise speed information with a good frequency response and thus has the potential for matching tyre and groundspeeds to a fraction of a knot.

Third, a system could be used to provide helicopters with a complete air data and three axis ground velocities (Fig 22). The air data is measured ahead of the rotor downwash, which is difficult to take into account using air data sensors on the fuselage. The horizontal ground velocities and climb rate can be measured using a vertical axis conical scan. These velocities can be used to establish a steady hover by maintaining zero velocity. The usual problems of identifying zero velocity with a Doppler system can be overcome with a laser.

Further developments of frequency modulated and pulsed Doppler CO_2 lasers will lead to systems which can measure range to solid objects as well as velocity components. Also it is reasonable to expect that the need for high pressure or liquid gas for cooling detectors will become a thing of the past with developments in thermo-electric cooling.

In conclusion, it is relevant to remember that most of the more sophisticated elements are common to the systems needed for all the tasks mentioned in this paper. Only the

scanning system with its demodulator and the final processing of the Doppler Spectra are specific to individual applications.

REFERENCES

(1) R. Callan, J. Cannell, R. Foord, R. Jones, J.M. Vaughan, D.V. Willetts and A.A. Woodfield: Proc. 5th National Quantum Electronics Conference, Hull 1981, Pub. J. Wiley & Sons Ltd., Chichester, England.

(2) R. Callan et al: RSRE Research Review No. 6, 1982, Paper 39.

(3) C. Bullock: Turbulence detection by laser. Interavia Magazine 1/1981, pp. 82-83.

(4) R. Foord, R. Jones, J.M. Vaughan and D.V. Willetts: Precise comparison of experimental and theoretical signal to noise ratios in CO_2 laser heterodyne systems. (To be published).

(5) Aircraft Accident Report - Pan American World Airways, Clipper 759, N4737, Boeing 727-235. New Orleans International Airport, Kenner, Louisiana, July 9 1982. US National Transportation Safety Board, NTSB-AAR-83-2, March 1983.

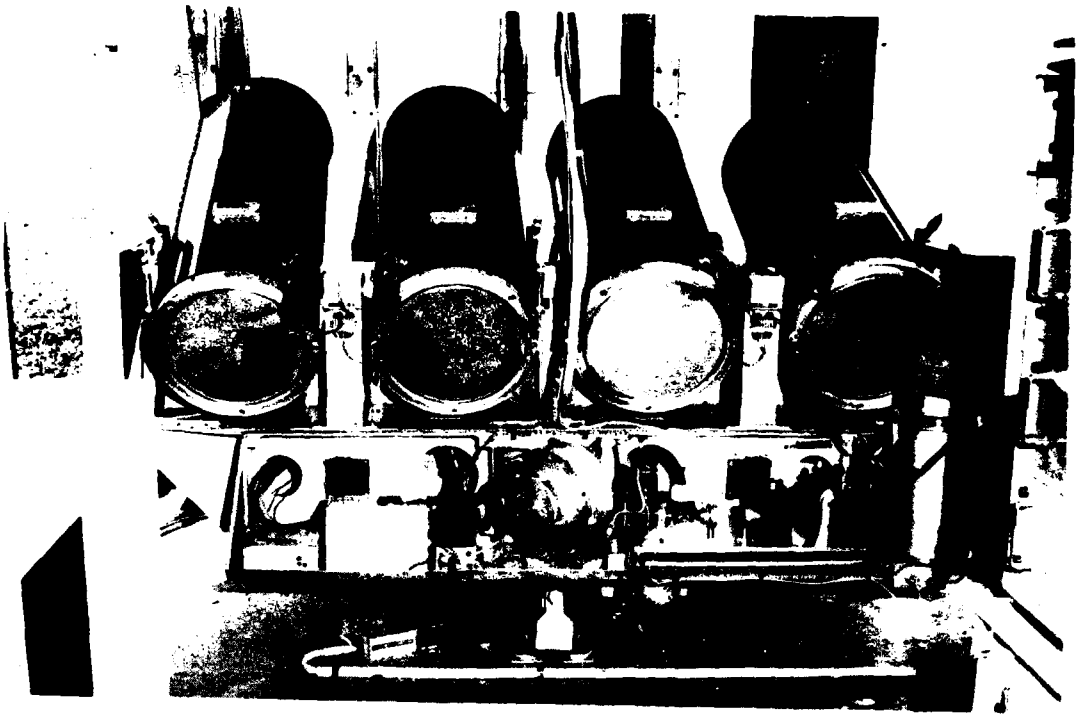


FIGURE 1. The Combined Wind And Vortex Experimental Range (CWAVER) equipment. In one mode of operation the outgoing laser beam was switched between the four telescopes in which the range and elevation could be rapidly altered.

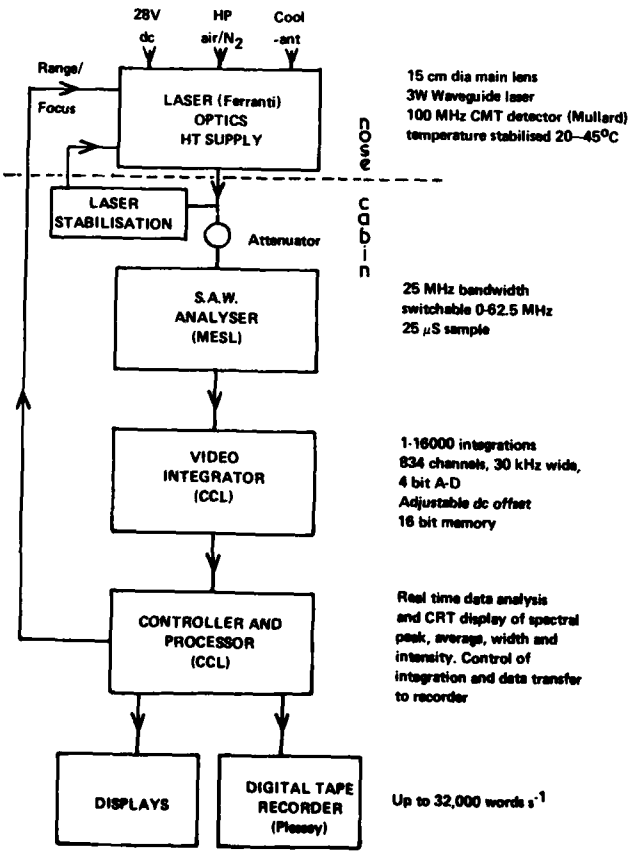


FIGURE 2. Block diagram of the Laser True Airspeed System (LATAS). The principal equipment parameters are shown and also (in brackets) the manufacturers.

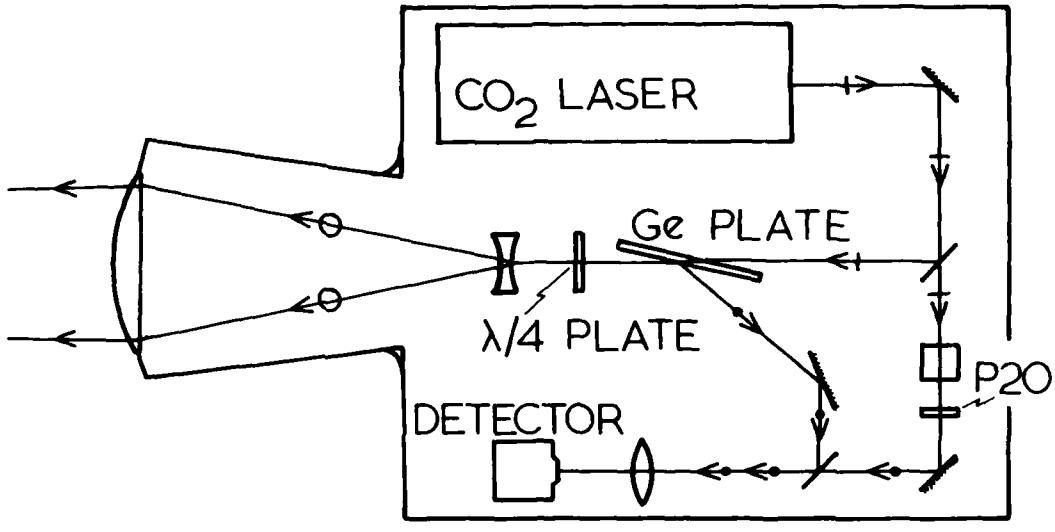


FIGURE 3. The layout of the Optics Head in the LATAS equipment. Polarisation techniques and germanium and quarter wave plates promote good efficiency.

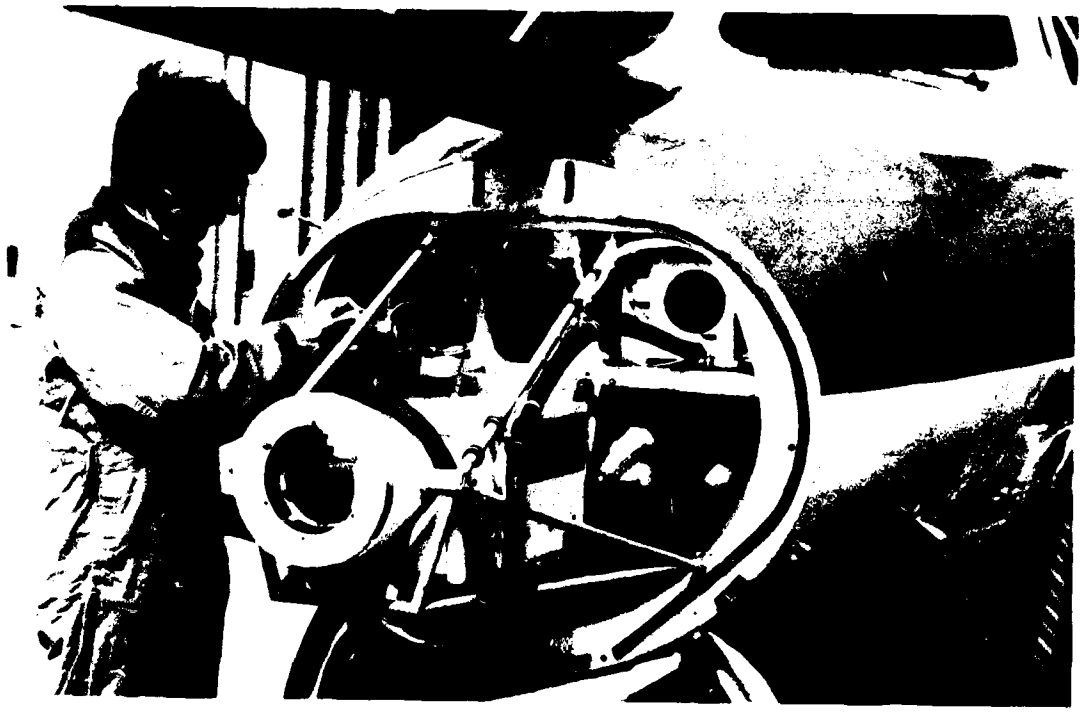


FIGURE 4. The Optics Head mounted in the nose of the HS125 trials aircraft of Flight Systems Department, RAE Bedford.

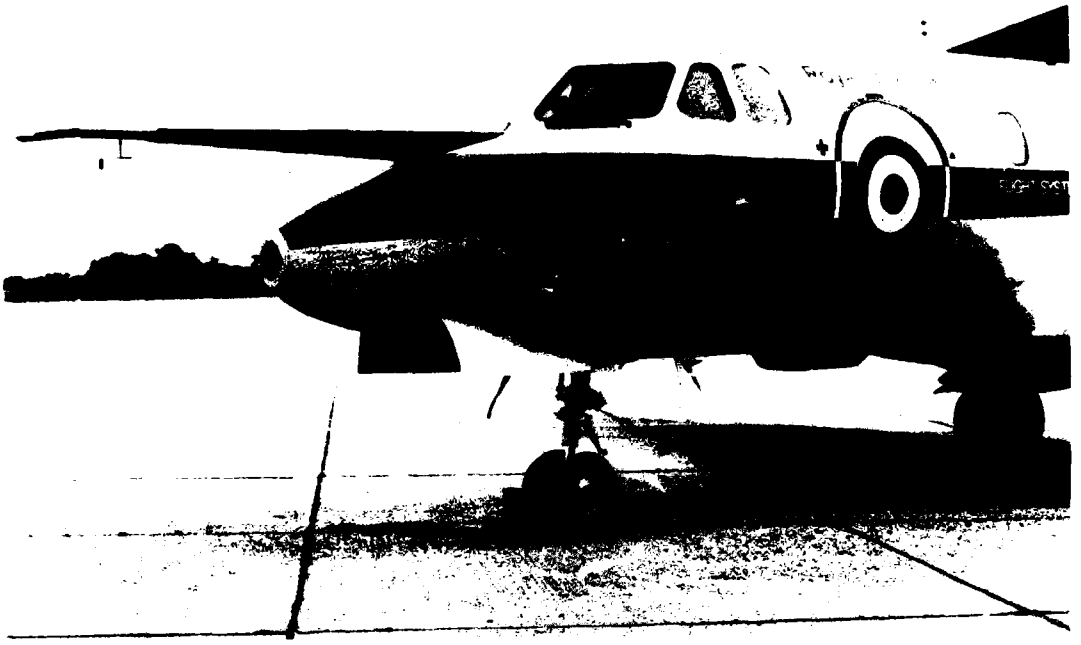


FIGURE 5. The HS125 aircraft showing the nose cone and 20 cm diameter germanium window. See also figure 8.

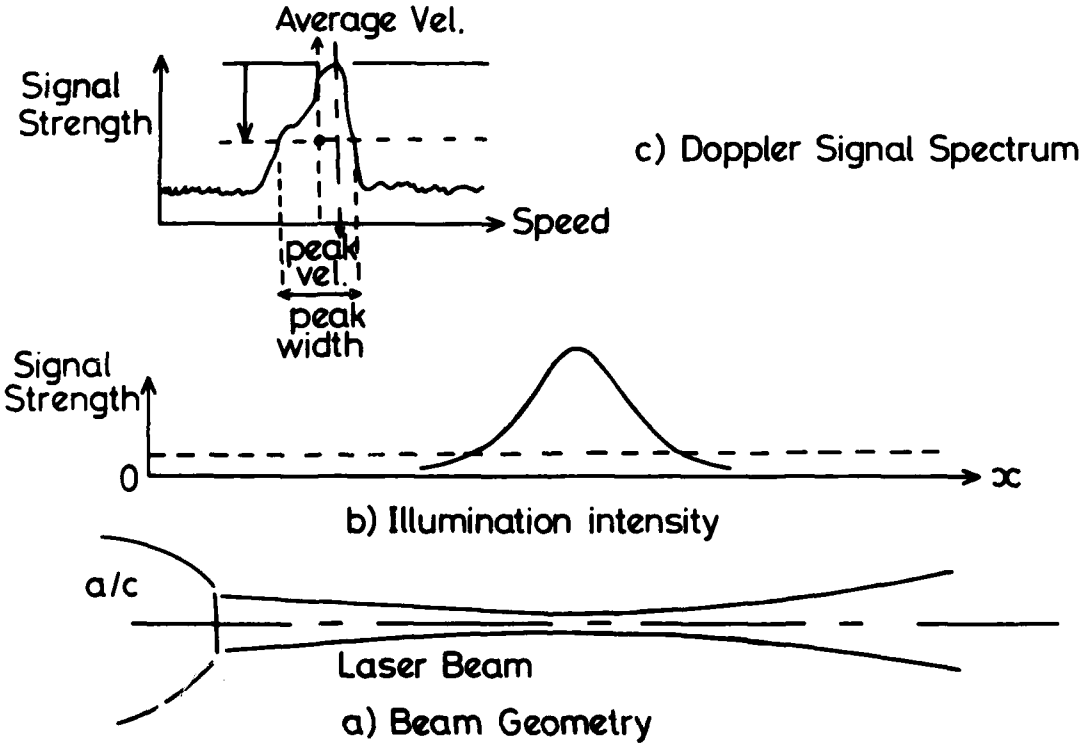


FIGURE 6. Parameters of a focussed CW laser anemometer showing schematically the laser beam, the illumination, and Doppler spectrum.

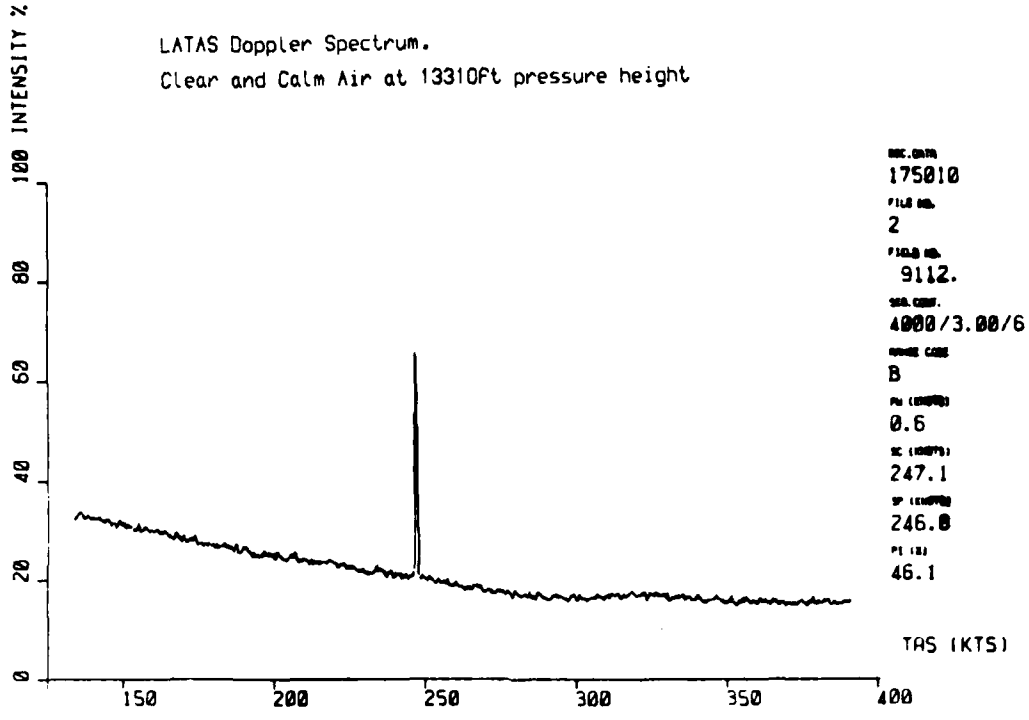


FIGURE 7. LATAS Doppler spectrum in clear and calm air at 13310 ft pressure height. Note the narrow Doppler signal with peak corresponding to a true airspeed of 246.8 knots.

FIGURE 8. The nose cone of the HS125 aircraft showing the germanium window and diamond-like hard coating (supplied by Barr and Stroud), after two years of flying. The nose cone shows considerable surface abrasion after about 30 flights since repainting.

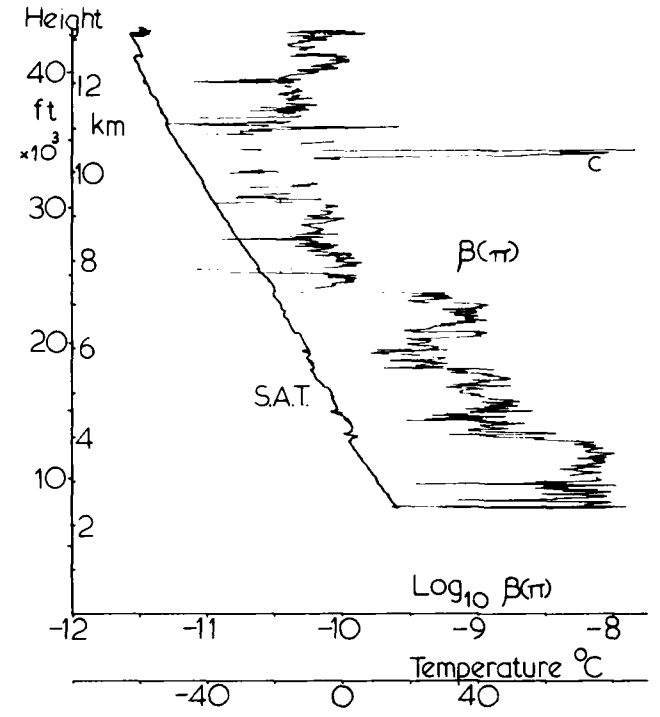


FIGURE 9. Atmospheric backscattering coefficient $\beta(\pi)$ sr^{-1} m^{-1} and static air temperature SAT versus indicated height in ft and km (ASL). The strong return at (c) at 33000 ft later in the day became a visible layer of cirrus cloud. [Flt 772, am 1 July 1982, from Jefferson County Airport Colorado, surface visibility > 80 miles].

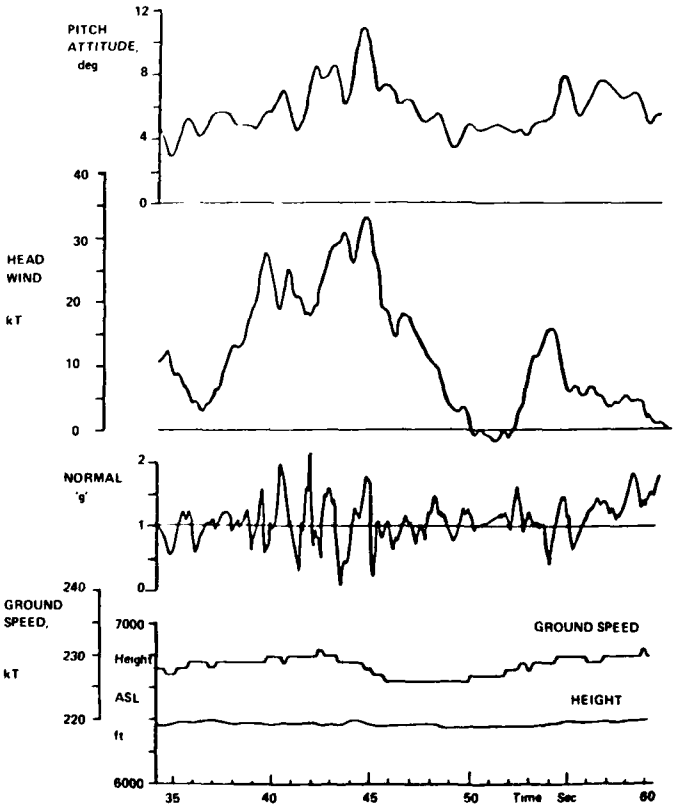


FIGURE 10. Passage through a thunderstorm microburst recorded at the JAWS project in Colorado [Flt 792, pm 14 July 1982, run 3]

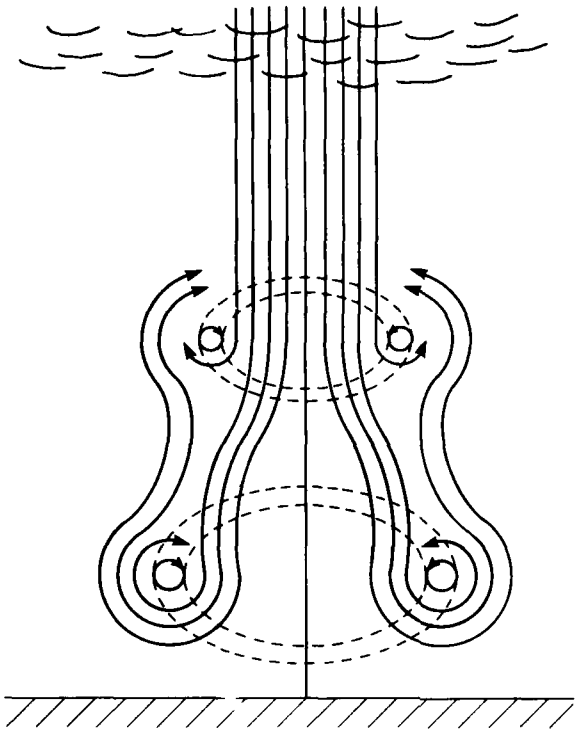


FIGURE 11. Vortex model of a microburst showing streamlines.

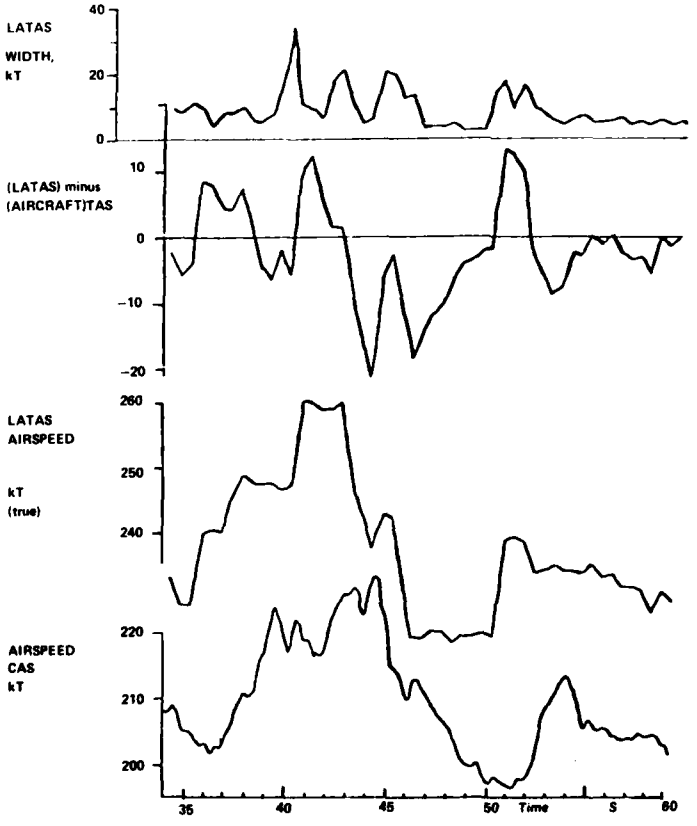


FIGURE 12. Passage through a thunderstorm microburst showing the response of the RAE/RSRE LATAS system. [Flt 792, pm 14 July 1982, run 3, JAWS project]. Note the 2 sec lead (corresponding to ~ 250m) of the LATAS airspeed.

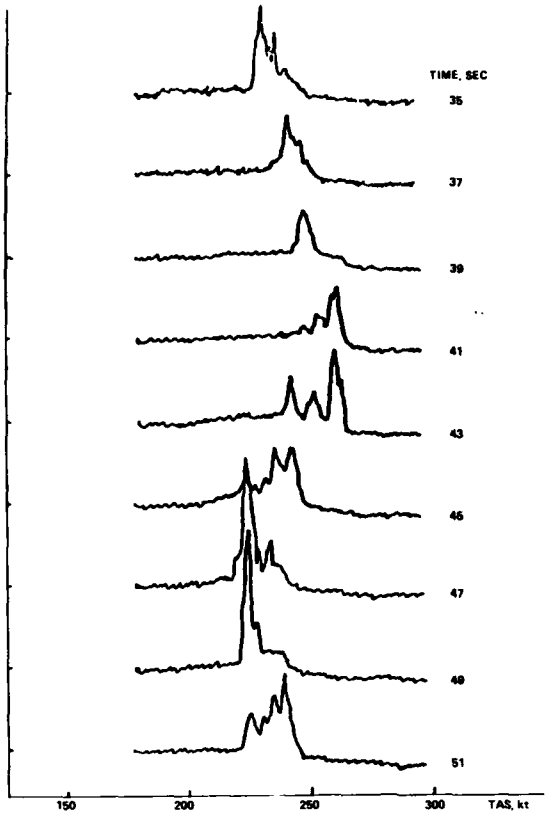


FIGURE 13. Successive LATAS Doppler spectra on passage through a thunderstorm microburst. The reference times are the same as for figures 10 and 12 [Flt 792, pm 14 July 1982, run 3].

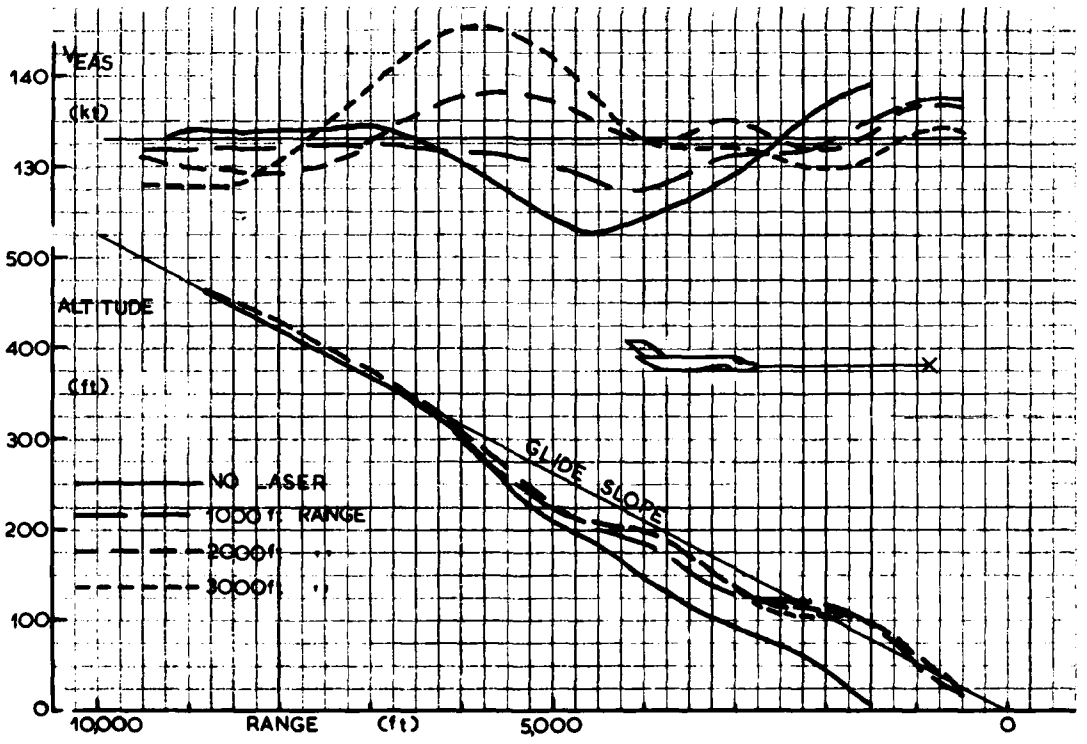


FIGURE 14. Flight simulations of a coupled approach through thunderstorm wind shear, for a BAC 111 medium jet transport, with and without LATAS type advance warning from different ranges. The thunderstorm cell was set at 4,500 ft range, and offset 1500 ft.

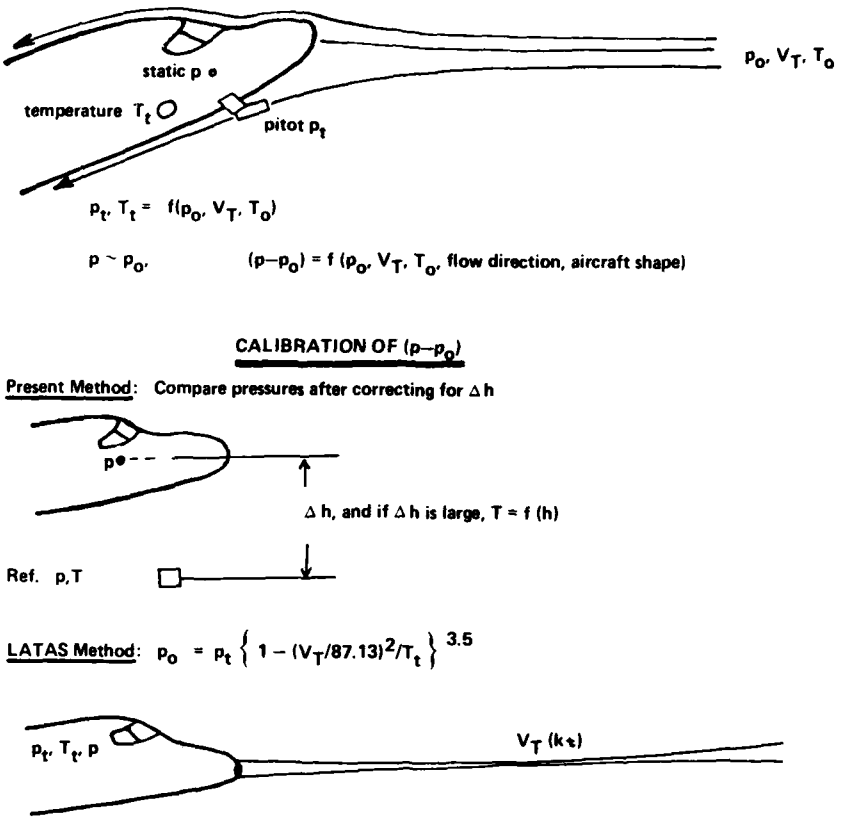


FIGURE 15. Calibration of pressure error ($p-p_o$); comparison of present methods and LATAS method

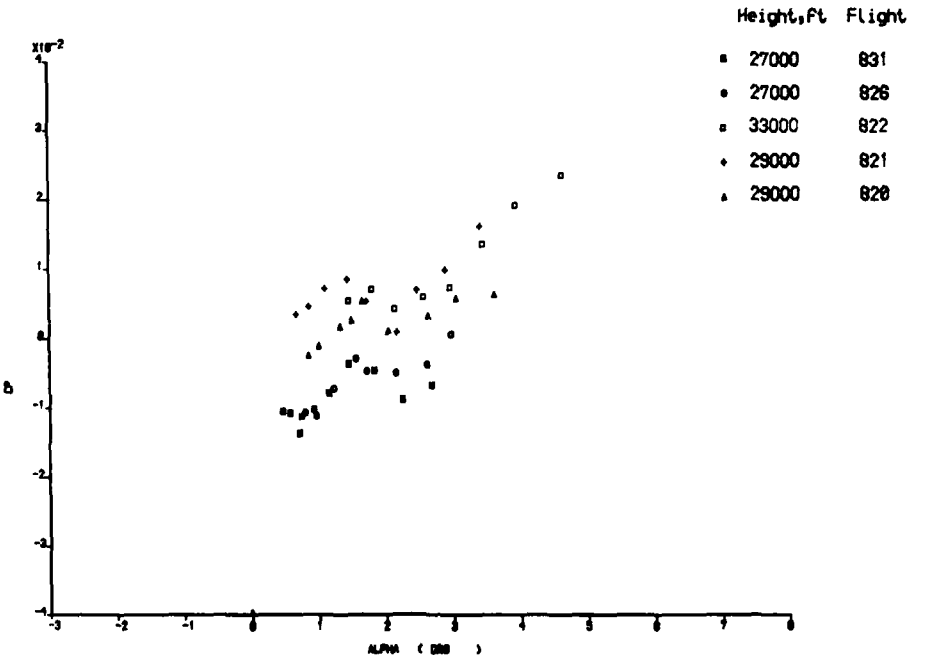


FIGURE 16. Pressure errors on the HS125 aircraft measured by LATAS.

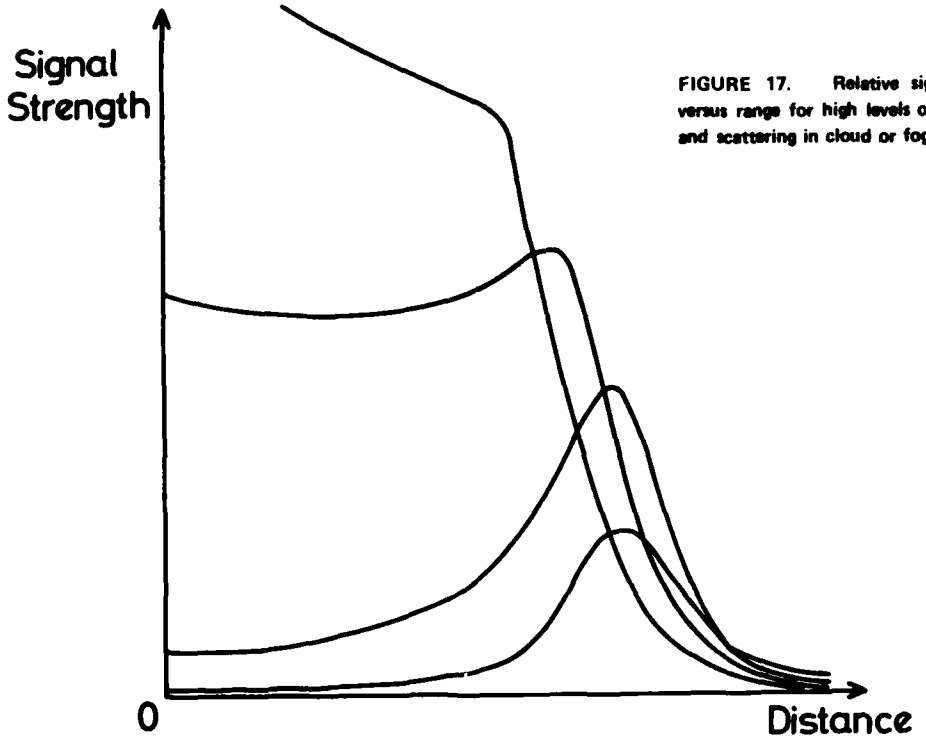


FIGURE 17. Relative signal strength versus range for high levels of attenuation and scattering in cloud or fog (schematic).

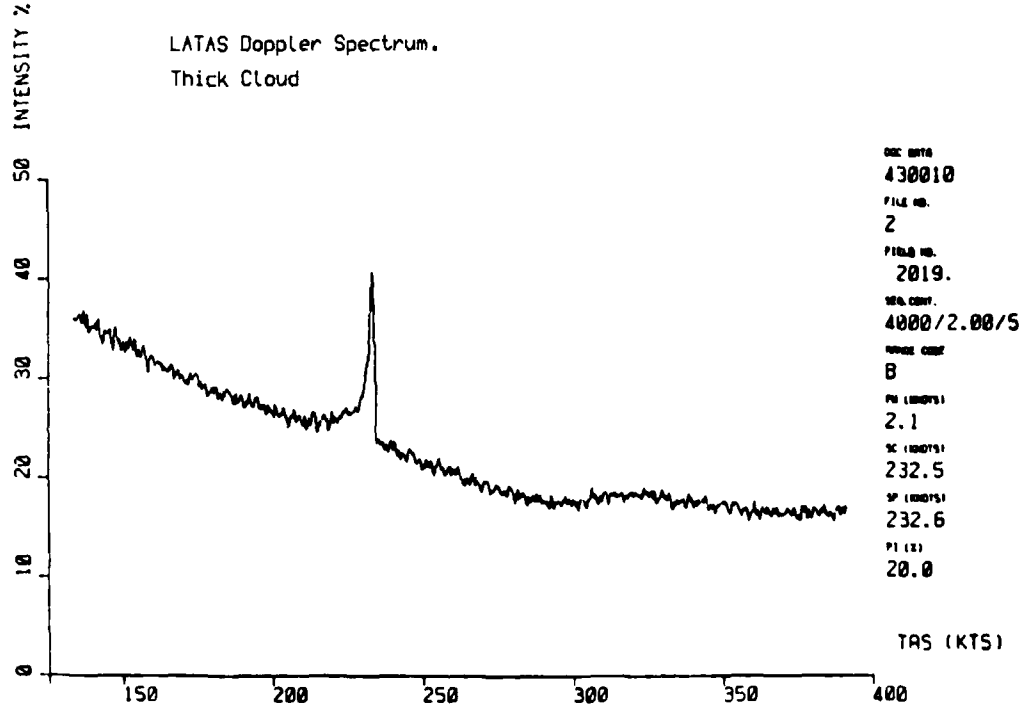


FIGURE 18. LATAS Doppler spectrum recorded in thick cloud.

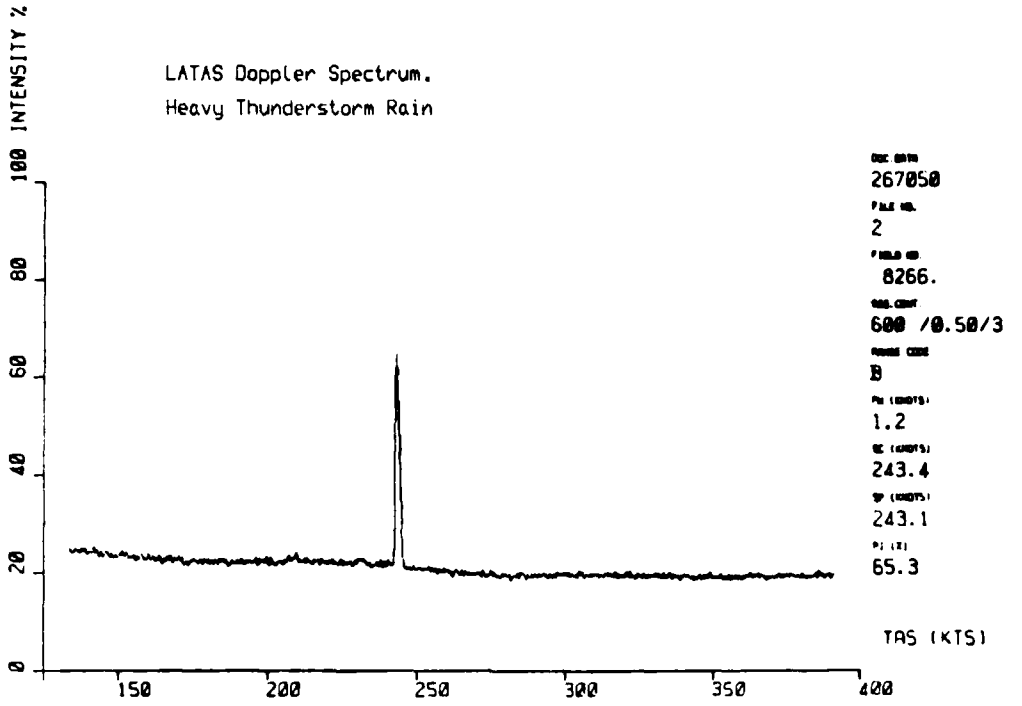


FIGURE 19. LATAS Doppler spectrum recorded in heavy thunderstorm rain.

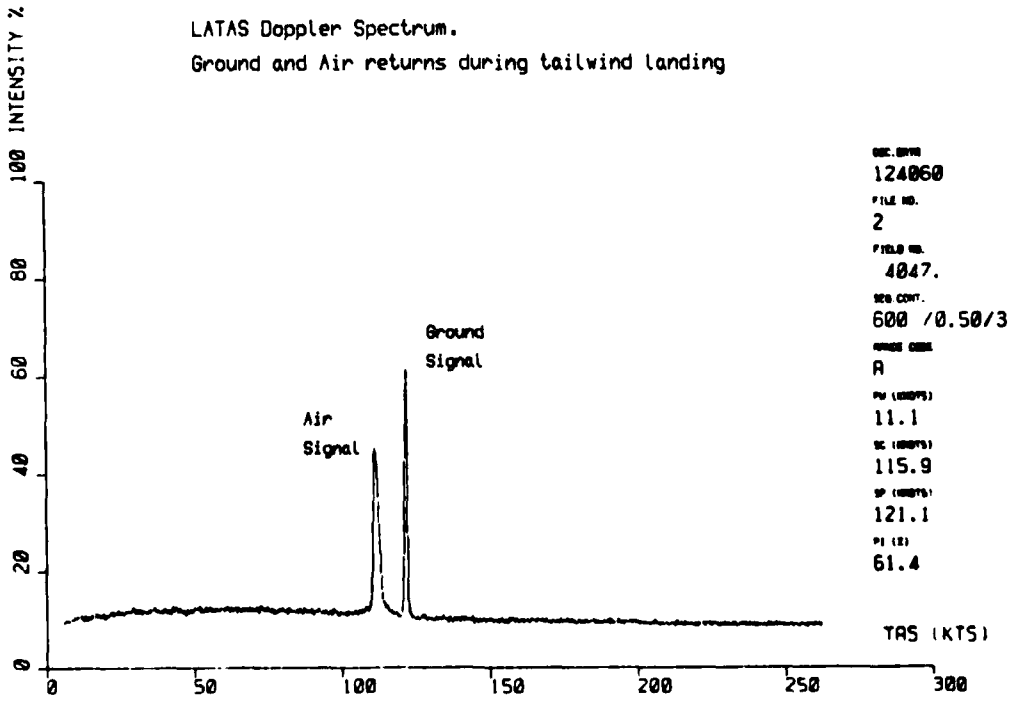


FIGURE 20. LATAS Doppler spectrum recorded with ground and air returns during a tailwind landing. Note the slightly broadened spectrum from turbulent air.

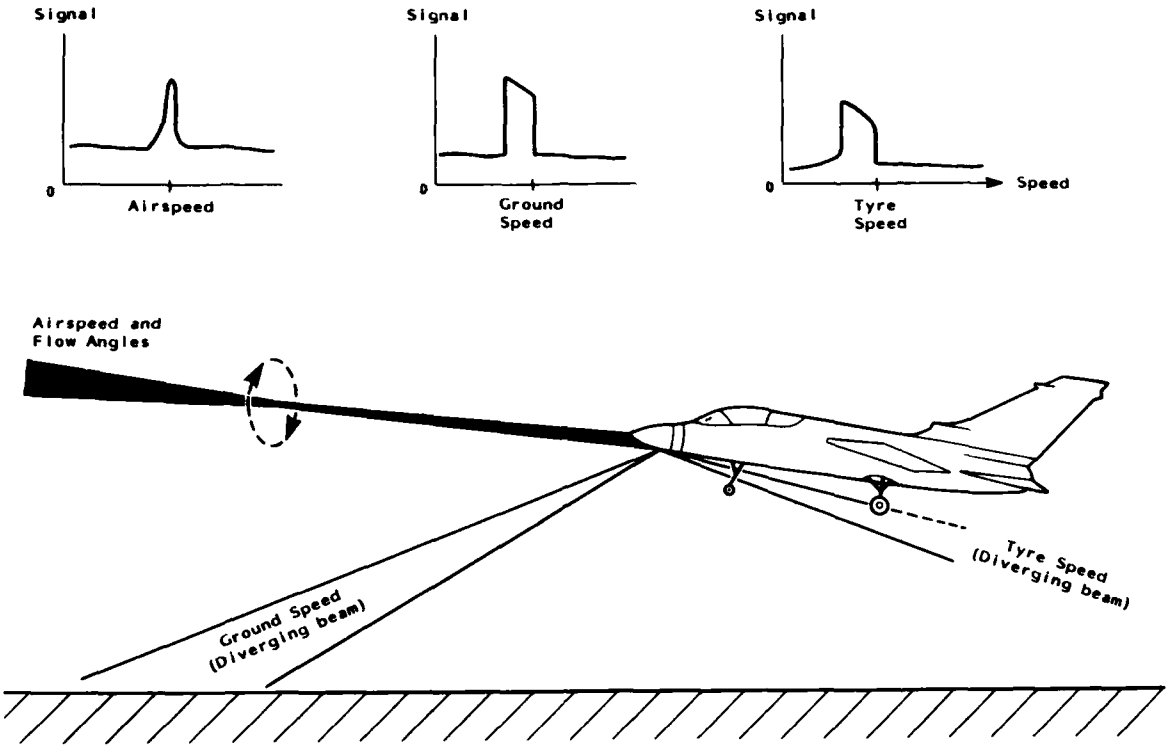


FIGURE 21. CW Laser Doppler system for airspeed, ground speed and tyre speed.

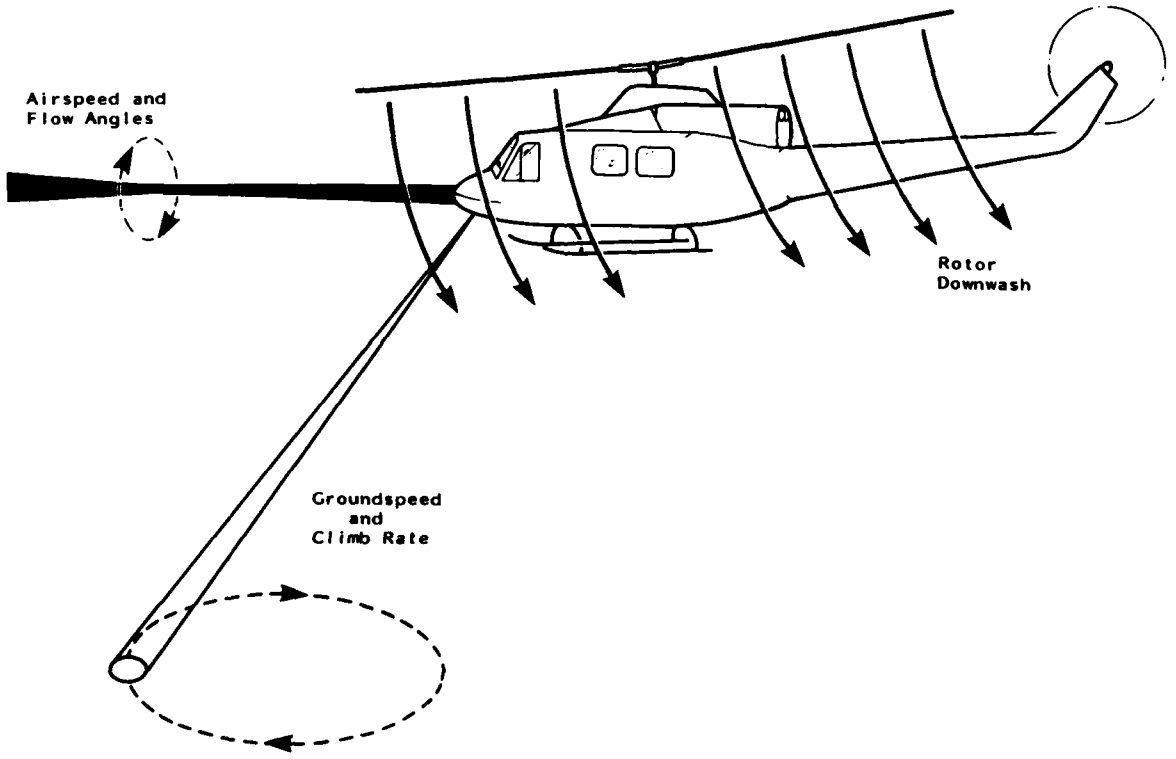


FIGURE 22. CW Laser Doppler system for air data, ground speed, climb rate and automatic hover on helicopters.

AUTOMATIC SPEECH RECOGNITION AS A COCKPIT INTERFACE

by

R. G. White
Royal Aircraft Establishment
Farnborough, Hants, UK

SUMMARY

This paper discusses the future use of automatic speech recognition machines for the management of cockpit systems and for low bandwidth air-ground communications. After describing the principles of speech recognition, the benefits afforded to the aircrew are considered. These are balanced against the difficulties of operating in the cockpit environment. The effects that this maturing technology is likely to have on cockpit design and operating procedures are discussed. The paper also considers techniques for training machines to recognise the aircrew's utterances. It is concluded that the future for the avionic applications of automatic speech recognition is promising. A great deal of relevant research is being undertaken within NATO countries, including flight research programmes planned for 1982.

1 INTRODUCTION

The use of speech as a man-machine interface is an old concept as those acquainted with the story of "Ali Baba and the forty thieves" will know*. However, developments in automatic speech recognition over recent years have enabled the concept to be realised. Progress can be attributed to two major influences. The first is the revolution in integrated circuit technology, which produced cheap, compact, high speed computers. The second is the interest in a wide range of commercially attractive applications. Examples are automatic telephone transactions, such as flight reservation or the ordering of goods, computer interfaces, automatic dictation, automatic speech translation, aids for the disabled and stock control. It will not be long before voice control of domestic appliances such as the television, electronic games or automatic garage doors is available. More recently there has been growing interest in the use of automatic speech recognition for military communications and aircraft cockpit management.

The severe cockpit environment has caused airborne applications to be slower in coming to fruition than their ground-based counterparts. Indeed, the author is unaware of any operational airborne uses of automatic speech recognition and the present paper is consequently speculative in many respects. However, the potential benefits are so great that considerable effort is being expended to overcome the environmental problems.

A complete solution to the technical problems would not be sufficient to guarantee the successful application of automatic speech recognition in the air. The human factors requirements must also be understood. After a brief introduction to the current state of the technology, this paper outlines the benefits and problems of airborne applications. The paper also discusses the human factors of integrating and operating a system in the cockpit. The bibliography is presented under three headings: Applications, Human factors, and Technology.

2 RECOGNITION TECHNIQUES

Speech recognition machines typically have three basic elements: a speech pre-processor which converts the speech signal into a computer compatible form, a scoring mechanism which measures how closely an utterance conforms to a reference template, and a logic unit which uses the scores to select the most likely message. The general arrangement is shown in Fig 1. Often the logic and match scoring units will interact so that time is not wasted producing scores for illegal words or combinations of words.

2.1 Pre-processing

The pre-processing can take a number of forms. For example, the time-varying power spectrum of the speech signal can be measured using a bank of filters. About 20 filters, spanning the speech spectrum, are used and the outputs digitised for subsequent computer processing. Typically the spectrum is averaged over a period of 10 to 20 ms and sampled 50 to 100 times per second, each sample being known as a 'frame'. The spectrum can also be derived directly from the speech signal using a Fast Fourier Transform.

Linear predictive coding[1] is an alternative technique which characterises the speech signal in terms of the parameters of a filter, the transfer function of which best matches the spectrum of the speech signal. Its advantage over the previous two methods is that it produces a smooth continuous spectrum.

---

* "... the robber chief, standing in front of a part of the rock that was as steep as a wall, called out in a loud voice, "Open Sesame!" With a noise like thunder the rock began to gape." From the stories of 'The Arabian Nights', as retold by Amabel Williams-Ellis.

Another method of characterising the speech signal is to measure the zero crossing rate of the signal in a number of broad bands to estimate the associated formant frequencies.

2.2 Matching

At present all commercially available speech recognisers use a technique called 'whole word pattern matching' to compare the pre-processed utterance with the reference templates. Differences between the power spectra of an utterance frame and a frame from a reference template are expressed quantitatively as a 'distance' measure. By adding together the distance measures for corresponding frames from the utterance and the template a cumulative distance measure can be produced which represents the dissimilarity between the utterance and the template.

The main difficulty of this approach is knowing which are the corresponding frames, since utterances of the same word by the same speaker will vary in duration and generally this variation will be non-linear. Fig 2 illustrates how normalising the length of the utterance with respect to the template can be used to deal with this problem. The string of frames which describes an utterance of the word 'helicopter' and the string describing a reference template are represented by spectrograms and arrayed against each other. Each spectrogram is a three dimensional plot of frequency versus time, versus energy level, the darkness of the points on the spectrogram representing energy level. The line A indicates the correspondence between frames if no time alignment is used. This is a poor method of comparison since the shorter utterance is clearly mis-aligned with the template. Line B represents linear time normalisation when the utterance is stretched uniformly in time to have the same duration as the template. This technique is used in a number of commercially available equipments but gives only limited recognition performance because it does not take account of the non-linear time variation of speech. The best match is achieved by optimal non-linear time-normalisation, represented by the line C. This can be defined as the alignment which gives the lowest cumulative distance measure. It is this measure which is adopted in the more advanced recognition systems, using dynamic programming[2] to find the optimal path.

As described so far the technique is suitable for an isolated word recogniser, which requires the speaker to leave a minimum pause between words, the word lengths being the times between the periods of silence. However, the technique can be extended for use in a connected word recogniser, which would permit the speaker to utter groups of words without a pause. In principle, dynamic programming would then be used to find the optimum alignment of the utterance with each possible sequence of templates. In practice this would take too long for real time operation. For example, if up to five connected words were permitted using a vocabulary of ten words then the utterance would have to be compared with 111, 110 possible template sequences. A variety of more efficient methods of using dynamic programming have therefore been developed for practical connected speech recognisers and these are described in the literature[3-6]. Connected speech recognisers have to contend with two major problems: co-articulation, where one word is run into the next, and the extraneous um's and er's that are an intrinsic component of unrehearsed speech.

Whole word pattern matching is attractive because it requires no knowledge of human speech perception. The whole utterance is used in the matching process and no attempt is made to reduce the data by isolating features relevant to speech perception. Because human speech communication has evolved over many thousands of years it can be argued that ultimately the best speech recognition machines will be based on the human perception of speech. However, insufficient is known about how humans perceive speech, so that the design of such a system would be speculative at present. Nonetheless systems have been developed with some success based on the extraction of linguistic features prior to the matching process[7,8]. The ultimate success of this approach may well depend on how relevant these features are to human speech perception.

A scheme developed at the National Physical Laboratory in the UK[7] converts the speech signal into a string of phonetic features, which are classified as particular fricative sounds, voiced sounds or periods of quiet. For example, in Ref 7 a typical utterance of the word six is described as: F3 180, V6 90, Q1 200, F3 150. This indicates that the word began with a fricative type 3 lasting 180 ms, followed by a vowel type 6 lasting 90 ms, a quiet period type 1 for 200 ms and finally a fricative type 3 lasting 150 ms. The matching process compares the utterance string with the strings representing the reference vocabulary. A mis-match score is derived by adding up penalty values for different types of discrepancy between the utterance and template strings. Ref 8 describes a scheme where the matching process involves searching for phonetic features within the syllabic structure of an utterance. Both schemes are described as continuous speech recognisers, where words in the legal vocabulary can be extracted from continuous speech, without constraints on the number of words in a phrase. The schemes are well-suited to continuous speech recognition because they operate at phonetic feature level and therefore require no knowledge of word boundaries.

2.3 Decision logic

The logic unit has to decide whether the best matching score for an utterance is above the threshold set for recognition and whether the difference between this score and those for alternative words is sufficiently great to interpret the utterance according to the best score. Thus the threshold levels determine the recognition integrity. If they are set too low, too many erroneous recognitions will occur. However, if they are too high the speaker will find the system annoying to use because he will have to repeat utterances too many times before an interpretation is produced.

For a given level of confusibility between words in a vocabulary, the recognition accuracy can be increased by reducing the size of the vocabulary. In general, only certain combinations of words from the total vocabulary will have a meaning. Thus, in most applications the overall vocabulary can be divided into sub-vocabularies by using syntactical rules in the decision logic. By setting appropriate recognition thresholds the resulting improvement in recognition performance can be harnessed as a higher integrity or a higher recognition rate. A number of commercially available speech recognisers can be programmed to apply syntax in their decision logic. The cockpit applications of syntax will be considered in section 6.2.

The decision logic can also be used for error correction[7]. When the speaker erases a word which was uttered in error or misinterpreted, the logic can weight the first interpretation as an unlikely interpretation of the repeat utterance.

3 THE BENEFITS OF AUTOMATIC SPEECH RECOGNITION

Fig 3 shows the traditional representation of the pilot and aircraft as a closed loop control system. The cockpit, represented by the pilot, the displays and the controls can be considered as a two-way interface between aircrew and aircraft systems. Displays provide information for the man: controls enable the man to control the aircraft and its systems. In principle, any of the cockpit controls could be replaced by an automatic speech recognition machine, which would enable the pilot to control aircraft systems directly, by means of his voice - a process now commonly referred to as direct voice input or simply DVI. In practice, DVI will only be employed where it offers specific benefits to the pilot.

3.1 Reduced task loading

The modern cockpit is characterised by an increasing number of systems with which the pilot is expected to interact. The resulting demand upon the pilot's attention can be so high that the success of a mission is limited by the capacity of the human being to cope.

During an experiment conducted at the Royal Aircraft Establishment in UK[9] subjects were required to enter strings of data using either a keyboard or their own voice, whilst carrying out a tracking task. The speed of entry and the number of errors made were recorded. No speech recognition machine was involved and the voice inputs were recorded for subsequent analysis only. The only errors were those made by the subject; the simulated speech recognition system was assumed to be perfect.

The experiment showed that whilst the subjective difficulty of entering the data was similar for voice and keyboard, the voice inputs were quicker and more accurate. In an operational situation the pilot would spend less time entering data using DVI than if he used a keyboard and, provided the DVI recognition performance was sufficiently high, less time correcting errors. Thus the use of DVI for data entry would give the pilot more time to attend to the other aspects of the mission.

3.2 Head-out operation

DVI can confer spatial as well as temporal advantages. The most obvious benefit afforded by DVI is head-out operation. The physical constraints of current military cockpits require the pilot to devote his attention to areas inside the cockpit both frequently and for long periods, in phases of the mission when he has a conflicting requirement to look out. For example, changing radio channels during close formation in cloud poses a serious problem. The pilot of a following aircraft must take his eyes off the leader to look down at the radio control panel. Similarly the inflight entry of navigation data at low level is a dangerous in-cockpit distraction. Research has shown that the use of tactile devices for keying in such data can involve long periods of head-down operation[10,11,12]. The management of weapon systems and multi-mode radars also require the pilot to look inside the cockpit when he should be looking out. DVI can alleviate such problems, permitting the pilot to carry out essential operations at critical phases of flight, while maintaining visual contact with the outside scene.

3.3 Improved design of primary flight controls

During high 'g' manoeuvres accurate hand movements are difficult. The control of cockpit systems such as radar and electronic counter-measures (ECM), during aerial combat is inadequate, using some types and locations of tactile control. These functions would be better suited to DVI, provided that any problems of automatic speech recognition under 'g' do not outweigh those of tactile control.

The incentive to maximise head-out operation and to permit control of systems during high 'g' manoeuvres has led to the hands-on-throttle-and-stick philosophy (HOTAS)[13,14] in cockpit design. In some instances this has resulted in an excessive number of switches, buttons and thumb-wheels on the primary flight controls. By using direct voice commands for some of the HOTAS functions, the corresponding tactile controls could be relegated to panel positions, which would permit inceptor designs to be simplified. The re-positioning would be acceptable because the tactile controls would be reversionary. The ability to control cockpit systems while keeping both hands on the primary flight controls is important for helicopter operations, particularly during nap of the earth manoeuvring; DVI could offer great benefits.

3.4 Data bus compatibility

The DVI system will be a cockpit controller which translates the pilot's utterances into data and commands for transmission to a number of specific aircraft systems. Tactile control panels will also be available as alternatives to the DVI (see section 5.1). Maximum flexibility will be afforded if either method of control can take over from the other at any time. This will necessitate the sharing of information about the stage reached in the control procedure and will require two-way communication between the DVI system and the tactile controllers. The DVI could receive reference templates and test information from a 'plug-in' data store and could be required to transmit updated versions of this information back to the same store (see section 7.3). The system could be self monitoring with a built-in test (BITE) routine, to report its serviceability to the cockpit warning system. Such a two-way communication within a multiplicity of aircraft systems makes DVI ideally suited to currently favoured system architectures, based on data buses such as 1553 B.

3.5 Compatibility with multi-system control panels

Recent cockpit design philosophies have included a central control panel which, like DVI, can serve more than one aircraft system. The objective has been to maximise the common use of keys to save panel area and to permit more than one system to be controlled from the optimum panel position. The F-18 cockpit embodies this philosophy with an 'up-front' controller[14]. The procedures for using such panels will include a branching syntax which can be mimicked in the DVI system, making it possible for the pilot to alternate freely between the two methods of control.

3.6 Flexibility

The flexibility of DVI will permit extensions of DVI functions and relaxation of operating procedures to be achieved without physical changes to the cockpit. For example, radio channels will be selectable using the names of radio stations as an alternative to stud numbers or frequencies. The injection of letter-coded data will require an extended keyboard; the addition of the phonetic alphabet to the DVI vocabulary will not physically affect the cockpit controls. This particular facility will permit the direct use of UTM grid references for navigation.

When a keyboard is used to enter data, a single procedure must be followed rigidly. This permits the number of keys to be minimised. Thus to enter the longitude, "018 degrees 01 decimal 2 minutes east", the initial zeros are keyed in so that there is no need for 'degree' and 'decimal' keys; one of the digit keys doubles as an 'east' key. A DVI system could be arranged to accept the same information in a number of forms:

"... zero one eight zero one two east ..."
"... one eight degrees one decimal two east ..."
"... eighteen degrees one decimal two east...".

The incorporation of these alternatives into a DVI system will require any human factors advantages to be balanced against a loss of recognition performance due to the increased size of the sub-vocabularies.

3.7 Low bit-rate communications

The use of automatic speech recognition for very low bandwidth air to ground communications is being studied both in the UK[15] and USA. A bandwidth of about 3 kHz is required to transmit speech directly in real time, but a channel of this bandwidth is susceptible to jamming and interference. Using an automatic speech recogniser, it will be possible to transmit a code for each word in the message and consequently reduce the required bandwidth. For a vocabulary of several hundred words, each word code will require 10 bits and the bandwidth will need to handle about 30 bits/second for real-time transmission. The communications link can thus use a lower bandwidth giving increased immunity from jamming and interference[16]. Alternatively, the spare capacity of a higher bandwidth could be harnessed for additional coding, for secure communications or for faster than real-time transmissions. One example of faster than real-time transmissions makes use of radio-signal scatter from meteor trails for air to ground communications[17]. Here, messages are transmitted as a rapid burst during the fraction of a second taken for the meteorite to burn up in the atmosphere.

Automatic speech recognition will permit a lower bit-rate for transmission than will other low-bandwidth speech encoders. Vocoders, which code the time varying speech spectrum, require data transmission rates exceeding 1000 bits/second. A phoneme coder, which transmits codes representing speech sounds, will require about 200 bits/second. Neither of these devices recognises words and they can be used only to drive a speech synthesiser at the receiver to reconstruct the utterance, with considerable loss of speech quality.

Transmissions based on word recognition will drive a speech synthesiser to give a good quality rendering of each word in the vocabulary. It will also be possible to feed the word codes directly into a communications network such as JTIDS[15,18] (Joint Tactical Information Distribution System). Provided suitable limits can be placed on the range of valid messages, the coding could be made independent of the language spoken.

4 THE COCKPIT ENVIRONMENT

The benefits of automatic speech recognition have to be balanced against the difficulties of operating in the environment of the military cockpit, which is characterised by

poor voice communications, high levels of acoustic noise, vibration, high levels of 'g' in manoeuvring flight, high mental stress and fatigue. In general, utterances in the military cockpit will be subject to greater levels of noise and distortion than those associated with ground-based applications. An airborne speech recognition system will consequently be more complex in design and operation to meet similar recognition requirements.

The principal sources of continuous acoustic noise will be engines, aerodynamic noise and air conditioning. The level of such noise picked up by a mask microphone in a fixed-wing aircraft is sufficiently small[19] for effective speech recognition. The noise in helicopters is likely to be less acceptable because it is more structured and the open microphones used give less attenuation. An effective airborne speech recogniser must identify and quantify the noise content of the speech signal and allow for it in the recognition process. Noise can be measured by the direct application of a second microphone or by processing the signal from the pilot's microphone. During the intermittent noise of firing the aircraft guns the signal from the trigger should be used to dis-engage the speech recognition system.

Increasing acoustic noise causes the pilot to speak more loudly to maintain his ability to hear himself. Such increases in speech loudness are accompanied by spectral shifts, for which a simple automatic gain control cannot compensate fully.

Oxygen masks introduced the major problem of breath-noise due to gas flow from the regulator and through valves. One solution under investigation treats such noise as an utterance and isolates it by matching it to a breath-noise template.

The effects of the cockpit environment on the human operator can cause utterance distortion. High levels of 'g', vibration[20,21] and the use of oxygen safety pressure (when there is a possibility of atmospheric contamination) all affect the vocal tract directly. The speech recogniser must be sufficiently robust in its operation to function within the resulting speech variations. One approach will be to feed information concerning 'g' levels, vibration and mask pressure into the speech recognition system to facilitate a programmed response. Although these environmental effects are widely accepted as problems in automatic speech recognition, their effects on speech recognition performance have yet to be quantified.

The effects of stress and fatigue may be even more difficult to accommodate, and may require a system which can adapt to changes in the voice during the on-set of such phenomena. Although there is some understanding of the effects of stress[22] and fatigue on speech, their effects on recognition performance are still largely unknown.

5 EFFECTS ON COCKPIT DESIGN

5.1 Control panel compatibility

Early DVI systems will not be flown operationally without the back-up of tactile control panels. The panels will be driven in parallel with the DVI system so that independent of the control medium employed the status of the system will be indicated on the panel.

Current control panels for individual systems such as radio, TACAN and weapon selection are in general incompatible with DVI in this respect. The toggle switches and rotary knobs cannot be driven remotely by an electrical signal from the DVI system. Fig 4 shows a UHF control panel, which falls into this category. Fig 5 shows an alternative panel, based on a key pad and an electronic alpha-numeric display. In this design the control of volume and the mode selection is purely tactile. However, radio channels and frequencies can be selected by DVI.

In the situation depicted, channel 12 has been selected. Another channel number can be punched using the key pad or spoken by the pilot, if the DVI system has been engaged. In either case the channel will appear on the display as a one or two digit number preceded by the letters CH. If a frequency is to be set manually, the CH/FR button is pressed, or the word 'frequency' is spoken. The last manually-set frequency is displayed as a four figure number. The new frequency can be set manually using the key pad or verbally using DVI. If a further frequency is required this again can be set manually or verbally. If a channel is then to be selected the CH/FR button is pressed or the word 'channel' spoken. The last selected channel is displayed, and the new channel number entered as before. Following an error, the pilot would press the CLR button or say the word 'clear' to remove the digits from the display (or alternatively to re-instate the previous selection).

A separate button, labelled G is reserved for distress calls. On pressing this or speaking the word 'guard', the button is illuminated and the guard frequency 243.0 MHz appears on the display. On selecting another channel or frequency the guard frequency is automatically deselected and the button ceases to be illuminated.

Other DVI compatible equivalents to the UHF panel in Fig 5 are possible and other DVI vocabularies could have been chosen. This example illustrates the principles of designing and operating DVI compatible control panels, which can be applied to other systems. An early decision to make cockpit components compatible with DVI will save expensive redesigns and retrofits.

5.2 Information feed-back

The effect of operating any cockpit control is fed back to the pilot. In simplest form this can be the position of a switch, where the feed-back can be tactile or visual. The operation of flight controls results in a change of attitude, detected by viewing the outside scene or the attitude indicator, or by means of a stick shaker or an audio warning if the safe flight envelope of the aircraft is exceeded. Data selected using a keyboard is presented as a digital read-out, to be checked before being entered into one of the aircraft systems. DVI is no exception and the pilot must be able to confirm that his utterances have been interpreted and acted upon correctly.

Suitable feed-back media for DVI include synthetic speech and a range of display devices. The best choice will depend on the functions chosen for DVI and the tactical roles of the aircraft.

5.2.1 Synthetic speech

Synthetic speech is one means of providing feed-back for DVI. It does not compromise the 'head-up' benefits afforded by DVI and the combination of the two systems results in a natural dialogue. However, such an application of synthetic speech must be treated with caution.

The primary function of synthetic speech in the cockpit will be to provide specific warnings, *eg* height warnings or a replacement for the coded tones used to warn of missile attacks. The more general the applications of synthetic speech, the lower will be its effectiveness as a warning medium.

One crucial difference between audio and visual feed-back is that the former is transient and the retention of its message relies on the memory of the man, whilst the latter can be presented over a period of time to be accessed and re-accessed by the man without a memory load. Another difference is that the synthetic speech system determines the timing of the machine to man communications, *ie* when the machine speaks the man must listen. Visual feed-back allows attention to be allocated when convenient. These characteristics will allow the head-out benefits of synthetic speech to be realised for simple utterances, but visual feed-back will be more appropriate for entering strings of digits, *eg* radio frequencies and navigation data.

The importance of head-out operation will depend on the aircraft role. In air to air engagements, head-down visual feed-back will be unacceptable since this can result in the loss of visual contact with a target. However, in air to ground operations, where targets are relatively static and terrain and obstacle clearance are of prime importance, short glances inside the cockpit will be permissible. The frequency of outside world references then takes on a greater importance than their length.

5.2.2 Cockpit displays

The best visual feed-back for DVI is an electronic display, since this can offer a flexibility to match that of the DVI system. However, if the DVI were used to make a simple selection, the feed-back could take the form of a single lamp or an illuminated key. When used to select particular modes of the cockpit displays the appearance of the appropriate mode could be all the feed-back that was required.

An alpha-numeric read-out will be required to set up weapon parameters or navigation data. This can be provided by technologies, such as the CRT or flat panel displays based on liquid crystal, LED or plasma-panel technology. The flat panel devices can be of either matrix or dedicated alpha-numeric form.

The display technology may be less important than display position. Since DVI facilitates head-out operation, this must not be compromised by burying the feed-back display deep inside the cockpit. The importance of the display position will depend on the role of the aircraft. Visual feed-back on the HUD or on the helmet-mounted display (HMD) is a contender for the air defence role. However, this could clutter the pilot's forward view and deny display space to other important information. A dedicated display at, or approaching, coaming level will be a better alternative in some applications.

6 OPERATING PROCEDURES

6.1 Management of the DVI system

An acceptable DVI system will be automatic in its operation, with no pilot adjustments required. Only an on/off switch in the cockpit, to power or isolate the system, and a means of engaging and dis-engaging the system will be needed. Two alternative methods of engaging the system are proposed.

A press-to-engage button, dedicated to the DVI system can be mounted on the stick or throttle. This would be used in a manner similar to the press-to-transmit button used for the communications system. Such a scheme would have the advantage that the DVI system is only engaged when the pilot wishes to use it. However, this entails an additional switch on the primary flight controls, which must be justified by the benefits of DVI and, in particular, by reductions in other HOTAS functions.

An alternative is to arrange for the DVI to 'listen out' for a distinctive keyword to activate the system. This is within the capabilities of available technology. The DVI could be disengaged by a second command word, either the final word in a sequence of the syntax or a separate word, which could be spoken at any time. Unfortunately this scheme would allow the DVI to be inadvertently left engaged, when it would be subjected to inadvertent utterances by the pilot. Communications would pose no problems because the radio press-to-transmit button would override the active state of the DVI.

6.2 Syntax

A DVI system controlling a number of cockpit systems may require a vocabulary as high as 100 words, depending on the number and complexity of the systems being controlled and whether or not the phonetic alphabet is required in addition to the digits 0 to 9. The vocabulary required for an effective low bit-rate communications link will be between 500 and 1000 words. Section 2.3 described how syntax can improve recognition accuracy by breaking down such vocabularies into smaller groups.

The management of cockpit systems usually follow set procedures, which can be translated into a set of syntactical rules for use in the DVI decision logic. For example, Fig 6 shows the syntax for parallel voice and panel control of the UHF system described in section 5.1. The solid lines correspond to the syntax requirements when only the UHF system is controlled by DVI. The broken lines indicate the additional functions required to control a number of cockpit systems. The inertial navigation system is a strong contender for DVI control and a syntax for its management is presented in Fig 7. The facilities included are position fixing, changing destination and entry of navigation information for way point and off-set locations. The paths which pass along a single line of nodes are not shown in detail, but examples of the required word sequences are given.

The word CLEAR is used to retreat one or more steps in the syntax, the number of steps being defined by the syntax at each node. This permits errors by the man or the DVI system to be corrected. The word RESTART is used to return to the beginning of the syntax sequence so that other cockpit systems can be addressed. RESTART is also used for error correction, when it is more convenient to start again than to retrace a path through the syntax by repeating the word CLEAR.

DVI operation of the UHF system will require a total vocabulary of 17 words. The sub-vocabulary size or branching factor will vary between 5 and 12 with an average of 10. To use DVI to operate both the UHF and navigation systems, 64 words will be required. The branching factor will vary between 2 and 26, with an average of about 10. The total vocabulary size will increase further with the number of cockpit functions controlled by DVI. However, the distribution of branching factors is unlikely to change drastically. The syntactical structures illustrated in Figs 6 and 7 assume a mechanical switch to engage and disengage the DVI system. However, keywords could be used, as described in section 6.1. This would add 1 to the branching factor at each node from which the DVI system could be disengaged.

When a pilot uses DVI to control cockpit systems his utterances fall into two categories; information and executive commands. For example, storing the geographical coordinates of a way point can require a UTM grid reference to be set up on an opticator, and, when checked, entered into the inertial navigation system. The two functions have different recognition requirements.

On the one hand the recognition accuracy required to set up the coordinates is not critical since mistakes by the DVI or indeed the pilot can be detected and corrected. Operating with the large branching factors of the digit set or the phonetic alphabet can be practical. Ideally, for this function, the decision logic will set recognition thresholds which minimise the time spent by the pilot in correcting errors and repeating utterances for which no interpretation was produced.

On the other hand the verbal command to insert or not to insert the grid reference must be a high integrity operation because an error could result in erroneous coordinates being stored for the way point. For such executive functions, the branching factor must be minimised; the sub-vocabulary must contain no confusible words and the recognition thresholds must be set sufficiently high to give the required recognition integrity.

Low bit-rate communication poses a more difficult problem for automatic speech recognition than does cockpit management. There will still be limits on which sequences of words constitute a meaningful message, so that syntax can be used in the decision logic. However, the syntax is likely to be more complicated and to have much larger branching factors than a syntax for cockpit management.

6.3 Vocabulary

The vocabulary will be central to the operation of a DVI system and the choice of some words will be a compromise between the conflicting requirements of human factors and recognition system performance. The former will favour words which are easy to remember, which are directly related to the function performed and do not depart from established aviation jargon. The latter will require the words of each sub-vocabulary to be easily distinguished. In many cases, words in current aviation use have been chosen to avoid confusion during radio transmitted conversation and these should also be suitable for DVI use. This is especially true of the phonetic alphabet and the exaggerated enunciation of the numbers 5 and 9 as FIFE and NINAH.

The choice of vocabulary for the DVI syntax described in the previous section was arbitrary because no appropriate NATO standard existed. However, developments in automatic speech recognition are progressing so rapidly that the early entry of a DVI system into service can be expected, even as early as 1985, as part of the mid-life update of a current NATO aircraft. There is therefore an urgent need for an agreed standard covering not only DVI but also low bit-rate communications, so that diverse vocabularies are not adopted by different NATO countries by default. There are already a number of NATO standards dealing with military terminology and a standard for automatic speech recognition could be based on these. For example, STANAGs 3596, 3277 and 3377 deal with air reconnaissance communications and AAP6 is the NATO glossary of terms and definitions for military use. In the UK, a study of suitable vocabularies is in hand at the Royal Aircraft Establishment.

7 TRAINING

An automatic speech recognition machine which can recognise a large number of words spoken by a large number of people, without the machine being trained for each speaker will have great commercial potential. However, such a capability would be a gross overkill for cockpit applications because most cockpit speech recognisers will be addressed by only a single speaker during any one mission. Even in multi-crew aircraft, the number of speakers is unlikely to exceed 2 or 3. Also, as the vocabulary is unlikely to exceed 100 words for DVI applications and 1000 words for low bit-rate communications, it will be practical to train an airborne system for each speaker using the whole vocabulary. The consequent relaxation of the specification is one of the rare advantages that airborne applications of automatic speech recognition have over their more general, ground-based counterparts.

7.1 Training procedures

Current commercially available speech recognisers are trained automatically. Each speaker utters the words of the vocabulary in a specific order and a specific number of times. The machine itself processes the utterances to produce word templates. Whether individual templates are stored for each utterance of a particular word or whether an 'averaged' template is used depends on individual technologies.

The number of training samples used is a trade-off between performance and training time and is application-dependent. The optimum is affected by the machine used, the size and confusibility of the vocabulary, the size and specialism of the user population and the environment in which the system is to be operated. Work at RAE[23] using a single speaker to operate a NEC DP100 machine under laboratory conditions found a significant improvement in the recognition of strings of 5 numerals on increasing the training samples up to about 5. However, little was gained by supplying further samples.

It is important for the training data to represent the utterances to be analysed in the aircraft. The equipment used to obtain the training data should be representative of the aircraft intercommunication system acoustics, the microphone and the electronics. Alternatively, the automatic speech recognition system must compensate for the characteristics of the airborne communications system, which will typically be based on a mask microphone in a fast jet aircraft, a noise cancelling boom microphone in a transport aircraft, or a throat microphone in a helicopter.

The pilot must be completely familiar with using automatic speech recognition equipment before the final training utterances are recorded. It has been observed that inexperienced subjects often speak differently during training and control sessions. Training utterances can be spoken self-consciously, while the following test utterances are spoken as a question or even with the subject shouting at the machine. Familiarisation allows consistency to be achieved. In addition to the machine learning to recognise an operator's utterances, the operator must develop the complementary skill of speaking consistently to the machine.

Training routines for low bit-rate air to ground communications pose other problems. Because the vocabularies are much larger than for DVI, unrepresentative templates can be obtained through loss of concentration by the subject. The tedium of the routine can be reduced by splitting it into several sessions. An alternative technique is to record utterances for some of the words only and to generate templates for other words from this information.

Aspects of the cockpit environment which affect speech were discussed in section 4. To be effective the recognition algorithms must either be insensitive to environmental changes or must monitor such changes and compensate for them. The former is difficult to achieve. The latter requires either detailed knowledge of the effects of environmental changes on speech or separate reference templates for each condition. In the second case, additional training utterances may be spoken, either under simulated conditions or directly in flight.

7.2 Storage of training data

Training the machine will take at least several minutes, with the technologies that will be available for airborne use. In the operational environment requiring immediate 'take-off', it will therefore be unacceptable for the pilot to train the machine every time he climbs into the cockpit. A more practical scheme will store training data from each member of the squadron on each aircraft, with the pilot identifying himself on entering the cockpit. Alternatively, each pilot can carry a recording of his

training data, which can be fed into any cockpit speech recogniser. This will be more acceptable operationally since it places fewer limitations on which aircraft a pilot may use. The training data could form an additional data segment on a portable data store, such as the Ferranti PODS system, which is currently being introduced to transfer navigation and tactical information.

The training data will be pre-processed and recorded as word templates. However, it is also feasible to record raw speech and programme the automatic speech recogniser to process this in a BITE routine.

7.3 Adaptive systems

The performance of airborne speech recognisers will be enhanced by an ability to add to or update the training data automatically, using utterances made in flight. Reference templates can then be based on a much wider experience of the speaker's voice than that obtained from the training samples. Training time can be saved by reducing the number of initial training samples. The templates, initially produced under artificial conditions will become more effective for airborne speech recognition. The machine will be able to adapt to gradual changes in the speaker's voice. Alternatively, a short term adaptive system might track changes in the voice, due to the onset of stress or fatigue.

Syntactical constraints will enable the recognition accuracy to be monitored, obviating the problem that incorrectly interpreted utterances will degrade the templates. By limiting the authority of the adaptive mechanism uncharacteristic utterances can also be prevented from degrading the templates. For example, a minimum standard of match with existing templates can be set, outside of which the adapting mechanism will ignore the utterance. For an adaptive system to be viable, the communication with the portable data store must be two-way so that the updated training data can be retrieved for use in subsequent sorties.

8 CONCLUSION AND PROGNOSIS

This paper has outlined the principles of automatic speech recognition and has discussed its prospective use in military aircraft. The speculative nature of parts of the paper indicate the amount of research still to be done before systems will be adequate for operational flight. However, the future of automatic speech recognition in the air, both for DVI and air to ground communications, is promising. A great deal of relevant research is taking place within NATO countries and this is being coordinated by a NATO research study group dealing with speech processing (AC/243 (Panel III/RSG 10)). Much of this research is concerned with the fundamental techniques of speech recognition and is more broadly based than avionic or even military applications. Nevertheless the success of future avionic applications is dependent on such work.

At present, at least seven automatic speech recognisers are available commercially[24]. In 1982 several more will emerge including a number designed specifically for human factors[25] and flight[26] research. A large number of experimental systems are also being developed and, with the flight testing[26-28] of some of these, the immediate future promises to be important in the development of automatic speech recognition as a cockpit interface. Flight trials will be essential to assess the balance between the benefits of automatic speech recognition and the problems of achieving an effective system in the airborne environment.

Acknowledgments

The author wishes to acknowledge the advice and information given by Mr J.S.Bridle of the Joint Speech Research Unit at Cheltenham and by Dr R.K. Moore of the Royal Signals and Radar Establishment at Malvern, who also supplied the spectrograms used in Fig 2.

REFERENCES

1 J.D. Markel, A.H. Gray: "Linear prediction of speech." Springer, 1976

2 T.K. Vintsyuk: "Speech recognition by dynamic programming methods." Kybernetika (Cybernetics), No.1, 1968

3 H. Sakoe: "Two-level DP matching - a dynamic programming-based pattern matching algorithm for connected word recognition." IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-27, No.6, pp.588-595, December 1979

4 J.S. Bridle, M.D. Brown, R.M. Chamberlain: "An algorithm for connected word recognition." JSRU Research Report 1010, October 1981

5 Cory S. Myers, Lawrence R. Rabiner: "Connected digit recognition using a level-building DTW algorithm." IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-29, No.3, June 1981

6 R.K. Moore: "Dynamic programming variations in automatic speech recognition." Proc. of Inst. of Acoustics

7 B.E. Pay, C.R. Evans: "An approach to the automatic recognition of speech." Int. J. Man-Machine Studies (1981), 14, 13-27

8 Timothy C. Diller: "A linguistic approach to speech recognition." CATC Electronic News, Summer 1981, pp.28-31

9 J. Laycock, J.B. Peckham: "Improved piloting performance whilst using direct voice input." RAE Technical Report 80019, February 1980

10 P. Beckett, M.R. Southworth: "Future cockpit displays." British Aerospace unpublished Report

11 R.G. White: "Problems of using a touch-sensing display to enter navigation data." Unpublished MOD(PE) Report

12 R. Seifert: "Analysis of pilot input concepts and devices into integrated display/ control systems." AGARD 32nd GCP Symposium, Stuttgart, 5-8 May 1981

13 D.E. Oldfield, L.T.J. Salmon, Wg Cdr E.J. Shaw: "Flight Systems Department design for an advanced military cockpit." Unpublished MOD(PE) Report

14 Eugene C. Adam: "F/A-18 Hornet crew station." AGARD 32nd GCP Symposium, Stuttgart, 5-8 May 1981

15 B. Burgess: "Trends in future airborne communication systems." Jour. Electronics and Power, pp.608-610, September 1981

16 B. Hillam, G.F.Gott: "An experimental evaluation of interleaved block coding in aeronautical HF channels." AGARD CP 239, Munich, June 1978

17 John D. Oetting: "An analysis of meteor burst communications for military applications." IEEE Trans. Comms., Vol. Com.-28, No.9, September 1980

18 J.P. Sputz: "Joint Tactical Information Distribution System (JTIDS)." Signal, pp.53-54, August 1981

19 K.R. Maslen: "The outputs of some mask microphones in operational flight." RAE Technical Memorandum FS 98, October 1976

20 Mary E. Johnson: "A preliminary of the effect of 2-20 Hz whole body vibration on spoken numbers." RAE Technical Memorandum FS 152, January 1978

21 C.W. Nixon, H.C. Sommer: "Influence of selected vibrations upon speech. Range 2 cps to 20 cps and random." AMRL-TDR-63-49, June 1963

22 J.B. Peckham: "A device for tracking the fundamental frequency of speech and its application in the assessment of strain in pilots and air traffic controllers." RAE Technical Report 79056, May 1979

23 C. Ellis: "The effect of training on the performance of a connected speech recogniser." RAE Technical Memorandum FS(F) 446, 1982

24 George R. Doddington, Thomas B. Schalk: "Speech recognition: turning theory to practice." IEEE Spectrum, September 1981, pp.26-32

25 J.B. Peckham: "Functional overview of an advance speech recognition system." IEE Colloquium, 1981/46 - The Computer Recognition of Speech, London, 18 May 1981

26 Philip J. Klass: "Voice control system tests scheduled." Aviation Week & Space Technology, 11 May 1981, pp.83, 85, 89, 93, 94

27 Flight International: "F-16 to fly with AFTI/voice-operated cockpit." 12 December 1981, p.1743

28 Interavia: Air letter No.9852, 9 October 1981, p.7

BIBLIOGRAPHY

A APPLICATIONS

B. Beck, E.P. Neuberg, D.C. Hodge: "An assessment of the technology of automatic speech recognition for military applications." IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-25, No.4, August 1977

Flt Lt J.M. Clifford, BSc, RAF: "The use of voice-interactive systems in the military aircraft cockpit." Royal Air Force College, Cranwell, Report 14 GD ASC/PROJ/7/81

G. Collin: "Aural commands tried in Mirage 2000 simulator." Air et Cosmos, 19 September 1981

J.R. Costet, J.M. Melocco: "Utilisation de la commande vocale à bord des aeroneufs de combat." AGARD Conference Proceedings 312, May 1981, Paper 15

D. Davenport: "Voice for command and control." Voice Data Entry Systems Application Conference, sponsored by Lockheed, 7-8 October 1981

E.B. Davies, J.B. Peckham: "New developments in cockpit human interfaces." RAE Technical Memorandum FS 412, June 1981

A.M. Godwin, Dr J.C. Ruth: "Voice command: the next threshold in cockpit operation." 4th Digital Avionics Systems Conference, St Louis, 17-19 November 1981

N.R. Ganguli: "Speech analysis for man-machine communication systems." Acoustica, August 1981, pp.329-334

S. Harris, N.E. Lane: "Voice controlled avionics: programs, progress and prognosis." AGARD Conference Proceedings 312, May 1981, Paper 14

T.B. Martin: "Practical applications of voice input to machine." Proc IEEE, Vol.64, 1976, pp.487-501

A.G. Mertz: "A preliminary investigation of aural input/output systems for in-flight information retrieval." Naval Postgraduate School, Monterey, CA 93940, September 1978

R.P. Plummer, C.R. Coler: "Speech as a pilot input medium." MIT Proceedings, 13th Conference on Manual Control, pp.460-462

J.L. Taggart, Jr., C.D. Wolfe: "Voice recognition as an input modality for the TACCO preflight data insertion task in the P-3C aircraft." Naval Postgraduate School, Monterey, CA 93940, March 1981

R. Turn, A. Hoffman, T. Lippiatt: "Military applications of speech understanding systems." Rand Corporation, AD 787-394, 1974

M.J. Underwood: "Some system aspects of using speech recognition." IEEE Colloquium, 17 May 1981

B HUMAN FACTORS

T.R. Addis: "Human factors in automatic speech recognition." Research and Advanced Development Centre, Technical Note 78/1, 16 December 1977

W.R. Franks, P. Allen, J. Soutendam, I. Taylor: "UNIGEN - Universal language of aviation." Aviation, Space and Environmental Medicine, Vol.51, April 1980, pp.339-343

O. Hållén: "Voice command compared to push-buttons for data input in a simulated flight task." Royal Institute of Technology (KTH), Stockholm, Aero Report FL 338, December 1977

J. Hart: "Some human factor considerations in the specification of a voice-interactive man/computer system for use in aircraft cockpits." MSc thesis, 1980

I. Kuroda, O. Fujiwara, N. Okamura, N. Utsuki: "Method of determining pilot stress through analysis of voice communication." Aviation, Space and Environmental Medicine, Vol.47, May 1976, pp.528-533

S.J. Mountford, R.A. North: "Voice entry for reducing pilot workload." Proceedings of the Human Factors Society, 24th annual meeting, 1980, pp.185-189

H. Mutschler: "Untersuchung der Spacheingabe zur Erhöhung der Beobachterleistung bei der RPV-Aufklärung." Fraunhofer-Institut für Informations und Datenverarbeitung, Karlsruhe, December 1980

H. Montague: "Voice control systems for airborne environments." Rome Air Development Center, TR-77-189, June 1977

A.V. Nikonov, F.A. Solodounik: "Features of the speech signal during the cumulative actions of coriolis accelerations." Voenno-Med.ZH (USSR), V.7, 1972, pp.78-81

R.J. Petersen, N. Lee, C. Meyn: "Vocabulary specification for automatic speech recognition in aircraft cockpits." Logicon Inc., San Diego, California, Final Report, September 1978 to June 1979

G.K. Pook, J.W. Armstrong: "Effect of operator mental loading on voice recognition system performance." Naval Postgraduate School, Monterey, CA 93940, Technical Report NPS 55-81-016, August 1981

A.B. Quartano: "Human verbal behaviour considerations in the design of voice actuated hardware systems, for P-3C aircraft." Naval Postgraduate School, Monterey, CA 93940

C TECHNOLOGY

W. Ainsworth, P.D. Green: "Current problems in automatic speech recognition." In pattern recognition: ideas into practice, Ed. by B.G. Batchelor, Plenum Press, New York, 1978

M. Blomberg, K. Elenius: "Försök med automatisk taligenkänning som hjälpmedel för flygplanspiloter." Royal Institute of Technology (KTH), Stockholm, Technical Report STL-TR-1977-2

P.A. Bowden: "A flexible real-time micro-programmed filterbank for speech recognition." IEEE Colloquium, 17 May 1981

J.S. Bridle: "Automatic speech recognition: an overview." IEEE Colloquium, 17 May 1981

J.S. Bridle, M.D. Brown: "Connected word recognition using whole word templates." Proceedings of the Institute of Acoustics, Autumn conference, 1979, A9, pp.25-28

P. Denes, M.V. Matthews: "Spoken digit recognition using time-frequency pattern matching." J. Acoust. Soc. Am. 32(11), 1960, pp.1450-1455

H. Dudley, S. Balashek: "Automatic recognition of phonetic patterns in speech." J. Acoust. Soc. Am. 30, 1958, p.721

D.B. Fry, P. Denes: "The solution of some fundamental problems in mechanical speech recognition." Language and Speech, 1958, Vol.1, p.35

P.D. Green, W.A. Ainsworth: "Development of a system for the automatic recognition of spoken basic English." Proceedings of the British Acoustical Society, Spring meeting, April 1973

P.D. Green, W.A. Ainsworth: "Towards the automatic recognition of spoken basic English." Machine Perception of Patterns and Picture; Conference, Series No.B, 1972, Institute of Physics, pp.161-168

D.R. Hill: "An abbreviated guide to planning speech interaction with machines: the state-of-the-art." Int. J. Man Machine Studies, 1972, 4, pp.383-410

D.R. Hill: "Man-machine interaction using speech." Advances in computers, Vol.11, Ed. by F.T. Alt and M. Rubinoff, Academic Press, New York, 1971

G. Kaplan: "Words into action. Advanced technology." IEEE Spectrum, June 1980, pp.22-26

W.A. Lea: "Trends in speech recognition." Ed. Prentice Hall Inc, 1980

S.E. Levinson, M.Y. Liberman: "Speech recognition by computer." Scientific American, April 1981

R. de Mori, L. Gilli, A.R. Meo: "A flexible real-time recogniser for spoken words for man-machine communication." Int. J. of Man-Machine Studies, October 1970, 2/4, 317-326

J.D. Mosker: "Clear speech: a strategem for improving radio communications and automatic speech recognition in noise." Aural communication in aviation. AGARD Conference Proceedings CP 311, June 1981

A. Newell: "Speech understanding systems." Carnegie-Mellon University, Pittsburgh, Pen., May 1971

K.W. Otten: "Approaches to the machine recognition of conversational speech." In: Advances in computers Vol.11; Ed. by F.L. Alt and M. Rubinoff, Academic Press, New York, 1971

Dr D.S. Pallett, J.P. Rigonati: "Steps towards meaningful performance specification for voice I/O systems." Voice data entry systems applications conference, sponsored by Lockheed, 7-8 October 1981

L.C.W. Pols: "Real time recognition of spoken words." IEEE Trans. Comput., Vol.C20, pp.972-978, September 1971

W. Rabiner, R.W. Schafer: "Digital processing of speech signals." Prentice Hall Inc., Engleswood Cliffs, NJ, 1978

R. Reddy: "Words into action 11: a task oriented system." IEEE Spectrum, June 1980, pp.26-28

H. Sakoe, S. Chiba: "Dynamic programming algorithm optimisation for spoken word recognition." IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-26, February 1978, pp.43-49

H. Sakoe: "Two-level DP matching - a dynamic programming-based pattern matching algorithm for connected word recognition." IEEE Trans. Acoustics, Speech and Signal Processing, Vol.ASSP-27, No.6, pp.588-595, December 1979

H. Sakoe: "USA Patent, 4,059,725. Automatic continuous speech recognition system employing dynamic programming." 22 November 1977

S. Tsuruta: "DP-100 voice recognition system achieves high efficiency." Jour. Electronic Engineering, July 1978 (139), pp.50-55

Z.M. Velichko, N.G. Zagoroyko: "Automatic recognition of 200 words." Int. J. Man-Machine Studies, 2(3), 1970, pp.222-234

G.M. Whité, R. Neely: "Speech recognition experiments with linear predication bandpass filtering and dynamic programming." IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-24, April 1976, pp.183-188

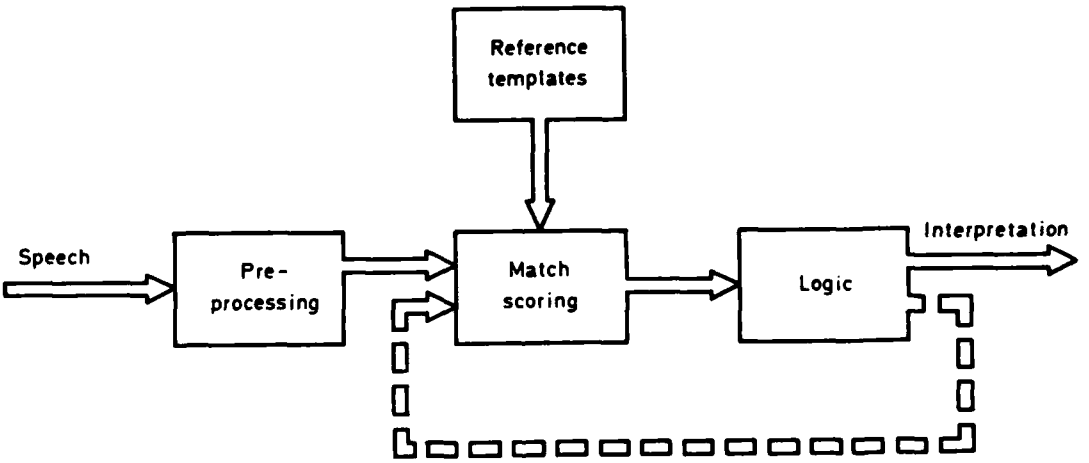


Fig 1 Automatic speech recognition system

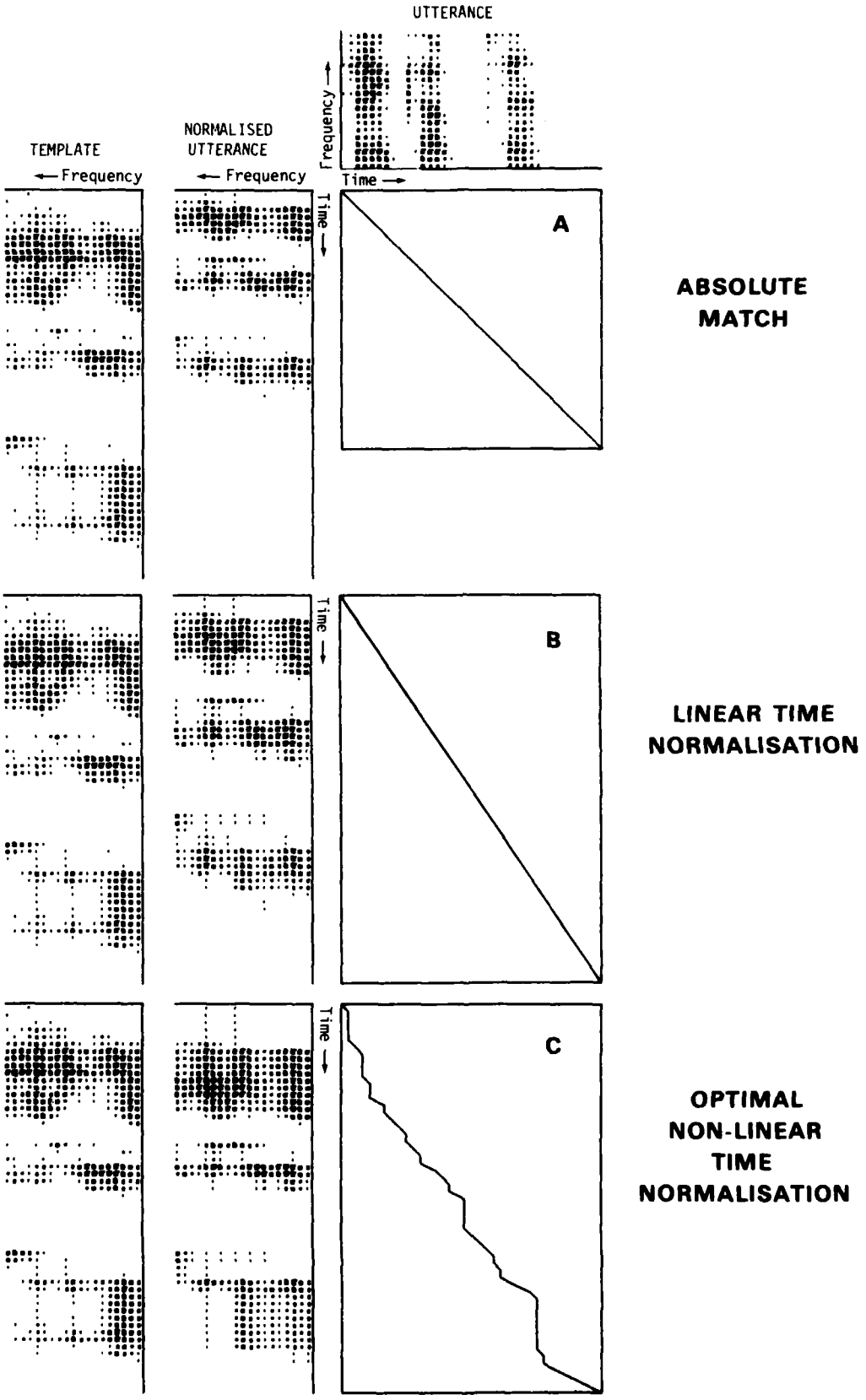


Fig 2 Speech pattern matching

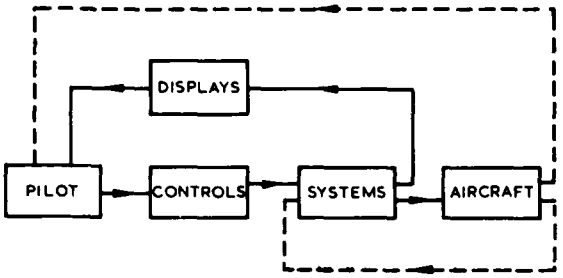


Fig 3 The cockpit as a man-machine interface

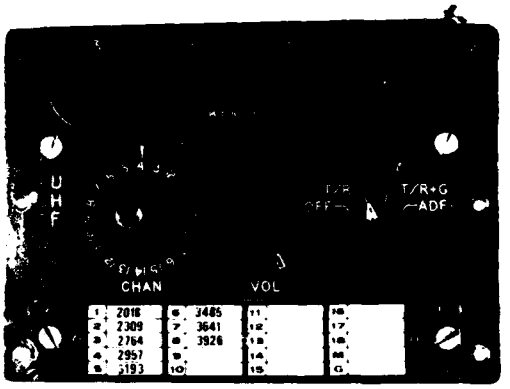
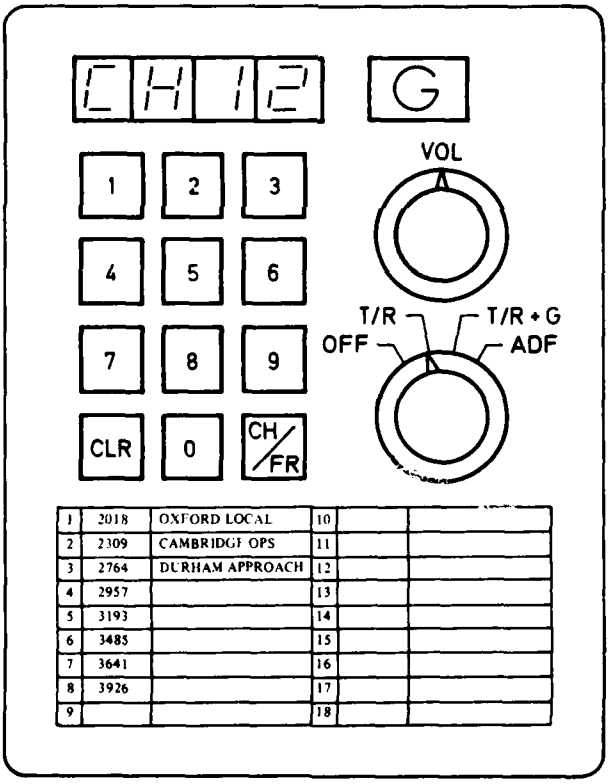
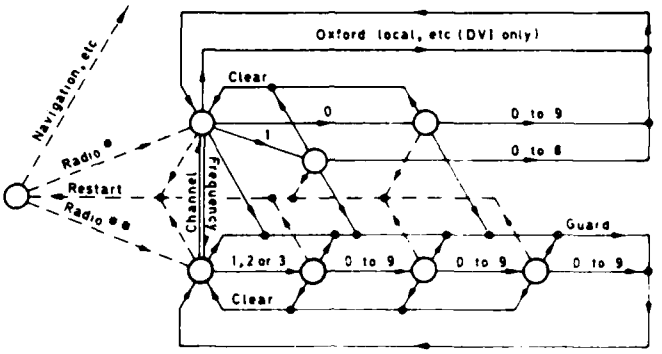


Fig 4 Conventional UHF controller



1	2018	OXFORD LOCAL	10		
2	2309	CAMBRIDGE OPS	11		
3	2764	DURHAM APPROACH	12		
4	2957		13		
5	3193		14		
6	3485		15		
7	3641		16		
8	3926		17		
9			18		

Fig 5 DVI compatible radio control panel



* If channel already selected

* * If frequency already selected

Fig 6 Syntax to control a UHF radio

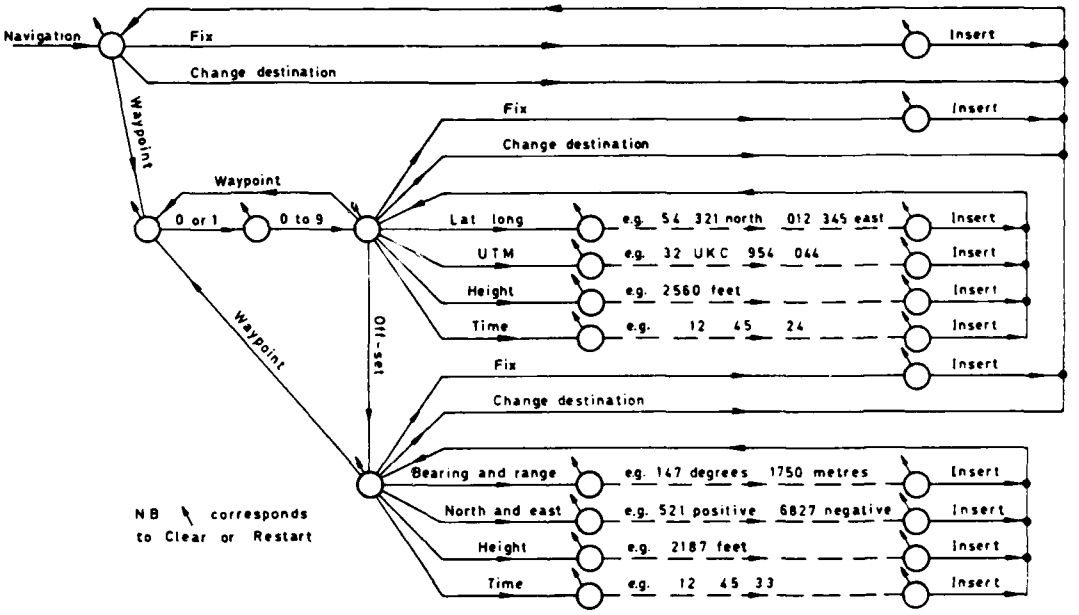


Fig 7 Syntax to control a navigation system

ANALYTIC REDUNDANCY MANAGEMENT FOR FLIGHT CONTROL SENSORS

by

James C. Deckert
The Charles Stark Draper Laboratory, Inc.
Cambridge, Massachusetts 02139
USA

and

Kenneth J. Szalai
National Aeronautics and Space Administration
Dryden Flight Research Facility
Edwards, California 93523
USA

SUMMARY

This paper reviews the formulation and flight test results of an algorithm to detect and isolate the first failure of any one of 12 duplex control sensor signals (24 in all) being monitored. The technique uses like-signal differences for fault detection while relying upon analytic redundancy relationships among unlike quantities to isolate the faulty sensor. The fault isolation logic utilizes the modified sequential probability ratio test, which explicitly accommodates the inevitable, irreducible low-frequency errors present in the analytic redundancy residuals. In addition, the algorithm uses sensor output selftest, which takes advantage of the duplex sensor structure by immediately removing a hard-failed sensor from control calculations and analytic redundancy relationships while awaiting a definitive fault isolation decision via analytic redundancy. This study represents a proof-of-concept demonstration of a methodology that can be applied to duplex or higher flight control sensor configurations and, in addition, can monitor the health of one simplex signal per analytic redundancy relationship.

1. INTRODUCTION

High levels of fault tolerance are mandatory for flight-crucial control systems upon which the safety of the vehicle depends. This is achieved in contemporary fly-by-wire control systems by providing redundant sets of computers, actuators, and sensors.[1-3] Under unfailed conditions, some means of device selection or consolidation is used to derive a nominal output. For example, in the case of triplicated sensors, a midvalue or averaging algorithm might be used to derive a single value to be used in control or navigation. In the case of actuators, one of the multiple channels might be selected for use with the others in a standby mode, or all channels may be active with force summing occurring on the output shaft.

Fault isolation and reconfiguration is accomplished by some means of comparison among like devices, with a failed device being detected by its deviation from the norm as established by a majority of other like devices. This approach must be abandoned when the number of devices reduces to two. Failures among one of two like devices cannot be isolated by cross-comparison. In many flight control systems, failures at the two-level result in the rejection of both remaining devices, and a backup mode of operation is established. For example, a failure occurring among two rate gyros might result in that feedback path being disabled altogether.

In order to improve the level of reliability of a redundant sensor subsystem, one or more of the following approaches must be pursued: increase the reliability of the existing sensors; add more like sensors; synthesize the needed sensor information through state reconstruction and estimation; or isolate the remaining good sensor should a failure occur at the two-level. This paper describes the formulation and flight investigation of an approach that accomplishes the latter objective, namely, the detection and isolation of a faulty sensor from a duplex sensor set. This general approach is not new. Various means of sensor self-check, such as rate gyro spin motor rotation detection and torqueing, have been used. The work described in this report, termed analytic redundancy management (ARM), utilizes sensor output signals exclusively. It grew from a desire to formulate an approach that would have a high level of coverage and integrity, make use of the large amount of sensor information inherent in an aircraft flight control computer, and that could exploit the computational capability inherent in digital flight control systems.

The technique described in this paper was developed for flight test evaluation on the NASA F-8 Digital Fly-by-Wire (DFBW) research aircraft. Algorithms were developed for 10 sets of redundant flight control instruments used on the F-8. The combination of theoretical and experimental work described in this paper demonstrates the validity of an approach that could be used to isolate faults at the duplex level or monitor a single sensor in a flight-critical use.[4]

The ARM technique is dual mode, with fault detection obtained by the comparison of like-sensor outputs, and fault isolation accomplished using modified sequential probability ratio tests (MSPRTs) operating on analytic redundancy residuals. This dual-mode

structure results in a low computational load in the normal no-fault situation, and it allows the MSPRTs to be made quite robust, since fidelity in the analytic redundancy relationships must be maintained for only the short time between detection and isolation.

Each MSPRT resembles a generalized likelihood ratio test[5] in which the failure time is known and the failure mode is assumed to be a bias of predefined magnitude, with the major difference being the inclusion of what is effectively a threshold offset in the MSPRT to accommodate low-frequency modeling errors and nominal sensor biases in the analytic redundancy residuals. The bias failure hypothesis is used not only because, in the absence of detailed failure mode information, bias failures are considered likely; but also because the resulting tests are quite effective in isolating other failures, such as ramps and scale factor errors, when observability is sufficiently high.

Sensor output selftest is utilized in the ARM algorithm to minimize the effect of a hard-failed sensor on the aircraft through the control system. (It should be noted, however, that because of its experimental nature the ARM algorithm is currently isolated from the F-8 DFBW aircraft control laws.) Four sensor signal status levels are defined in decreasing order of reliability: unfailed, provisionally failed, conditionally failed, and unconditionally failed. If both like signals have equal status, the average of the two signals is used in the analytic redundancy and control system calculations; otherwise, the sensor signal having the better status is used. An unfailed signal is declared provisionally failed when it differs from its value on the previous sample (and from the value of its companion signal if unfailed) by a predefined threshold magnitude. If this selftest violation disappears on either of the next two samples, the provisional failure status is removed. Conditional and unconditional failure status declarations are made by analytic redundancy MSPRTs as discussed below.

2. DETECTION AND ISOLATION TECHNIQUES

2.1 Fault Detection Using Direct Redundancy

For fault detection and isolation (FDI) purposes, it is convenient to define a faulty sensor as one having an output error magnitude larger than a stipulated bias failure magnitude (BFM), and in practice we would like to isolate an instrument having an output error magnitude of the order of BFM. Thus the BFM for each signal type is chosen to be larger than the observed output errors in good instruments and large enough to be isolated by the available analytic redundancy, and a signal fault is detected when the moving window average of the output of instrument one minus the output of instrument two is larger than three-quarters BFM in magnitude. This three-quarter BFM threshold results in equal probabilities of detecting a BFM/2 bias and not detecting a BFM bias. Stipulating these probabilities (10^{-4} for this study) and assuming white gaussian noise yields the required window size for each signal type as a function of BFM and noise variance.

2.2 Fault Isolation Using Analytic Redundancy

The SPRT[6] utilizes sequential observations of a process to decide which of two hypotheses concerning the probability distribution of the process is true. The SPRT is independent of the a priori probabilities of the two hypotheses, and minimizes the average number of observations necessary to reach a decision while meeting prespecified misclassification probabilities.[7] Because of these desirable characteristics and the simple form of the test, the SPRT is an ideal candidate for use with analytic redundancy for fault isolation. In particular, assuming that the process being observed is the difference between the output of one suspect sensor and a synthesized output using analytic redundancy, appealing choices for the two hypotheses are that the process has a mean equivalent to a BFM-sized bias (i.e., the sensor has failed) or that the process has zero mean (i.e., the instrument is unfailed).

Unfortunately, such factors as allowable biases on unfailed sensors, errors in the sensor input/output models, and parameter uncertainties in analytic redundancy relationships all contribute to low-frequency errors in analytic redundancy residual processes for unfailed sensors. Direct application of an SPRT to such a process may result in the acceptance of the failure hypothesis in spite of the fact that the sensor is operating within acceptable tolerances.

The MSPRT is a fault isolation test that systematically accommodates those irreducible factors contributing to low-frequency analytic redundancy residual errors that cannot be explicitly removed by modeling. The ARM algorithm utilizes the MSPRT to make conditional and unconditional failure status declarations following fault detection. Underlying the technique is the assumption that one of the following two hypotheses concerning the analytic redundancy residual process γ^j for suspect sensor j (j = 1 or 2) is true:

1) At time t_k, γ^j is gaussian with variance σ^2 and mean m_k (i.e., sensor j has failed)

2) At time t_k, γ^j is gaussian with variance σ^2 and mean 0 (i.e., sensor j has not failed)

Although a straightforward approach would be to design a test for each sensor that could accept either of these two hypotheses, and then declare a failure when either hypothesis

for either sensor was accepted, a more conservative approach is employed that avoids the tenuous situation of inferring that one sensor has failed merely because its companion sensor appears to be unfailed.

Consistent with the above discussion, following fault detection at time t_1 the MSPRT test statistic, the modified log likelihood ratio (MLLR), is defined at time t_n for suspect sensor j as:

$$u_n^j = \sum_{k=1}^{n} \left[\frac{m_k^j}{\sigma^2} \left(\frac{m_k^j}{2} - \gamma_k^j \right) + \frac{|m_k^j|}{\sigma^2} E_k \right] \tag{1}$$

and the following decision rule is used:

$$\begin{aligned} u_n^j &\leq \delta && \text{declare instrument j unconditionally failed, terminate the test} \\ \delta &< u_n^j < 0 && \text{declare instrument j conditionally failed, take another sample} \\ 0 &\leq u_n^j && \text{take another sample} \end{aligned} \tag{2}$$

In Eq. (1) the k subscript indicates evaluation at time t_k, and the mean m_k^j is computed assuming a BFM-sized bias in sensor j of sign consistent with the direct redundancy window at the detection time. The negative threshold δ in Eq. (2) is the original SPRT thresnold calculated using prespecified misclassification probabilities.

The last term in the summation in Eq. (1) differentiates the MLLR from the LLR of the standard SPRT, and represents the contribution of a postulated worst-case residual error magnitude at time t_k, E_k, to LLR calculation. Eqs. (1) and (2) indicate that the MSPRT is in essence a one-sided SPRT with a time-varying threshold offset arising from the worst-case error term. It follows that so long as the threshold offset is conservative, the misclassification probabilities for the MSPRT will be no larger than those specified to determine the original SPRT threshold.[8] The choice of the worst-case error magnitude for each analytic redundancy test requires considerable engineering judgment. An optimistic choice lowers the reliability of the test while an overly pessimistic choice may result in prohibitively long isolation times, although the inclusion of conditional failure declaration using a relaxed test criterion tends to lower the mean isolation times seen by the control system without corresponding increases in the ultimate misisolation probabilities. Specific choices for the worst-case error terms are discussed in subsequent sections on the particular forms of analytic redundancy utilized in this study.

In addition to the decision rules of Eq. (2), the ARM algorithm avoids open-ended isolation tests by utilizing an elapsed time limit (ETL) for each sensor type. If ETL is reached before an unconditional failure has been declared, the detected fault indication is removed, isolation computations cease, and the direct redundancy detection process is reinitiated. Because failure observability and worst-case error magnitude are often maneuver dependent, pilot response to notification that ETL for a sensor type has been reached could result in an enhanced fault isolation environment during the subsequent isolation period. Alternatively, reaching ETL could initiate hardware selfcheck routines.

3. DETAILS OF THE ALGORITHM

3.1 Overview

The ARM algorithm monitors ten duplex instruments aboard the F-8 DFBW aircraft: longitudinal accelerometer, lateral accelerometer, normal accelerometer, roll rate gyro, pitch rate gyro, yaw rate gyro, attitude platform, barometric altimeter, Mach meter and alpha vane. (Although the accelerometer and rate gyro complement is triplex, only a duplex subset is utilized by the ARM algorithm.) Additionally, a simplex beta vane is used in some calculations, but not monitored for failure. Each attitude platform gives outputs of Euler roll angle, ϕ, pitch angle, θ, and azimuth angle, ψ; and these outputs are considered to be three independent signal types although, in practice, the failure of one channel would probably dictate the failure of the unit.

From the large number of analytic redundancy relationships available, practical considerations and aircraft-specific signal-to-noise values reduce the number used in the ARM algorithm to four general types: rotational kinematics, altitude kinematics, translational kinematics and translational dynamics. In the following four sections, the analytic redundancy residual equations and worst-case error terms are discussed for each sensor type, grouped by the type of analytic redundancy employed. The lateral and normal accelerometers employ two types of analytic redundancy each while the remaining signal types employ one each.

To summarize the ARM FDI process, each signal type utilizes a threshold test on the moving window average of the difference in the duplex signals to detect a fault. Following fault detection, one MLLR is computed via Eq. (1) for each suspect sensor for each form of analytic redundancy used, and the MSPRT threshold logic of Eq. (2) is applied to the lowest MLLR. This process is repeated until an unconditional failure is declared or ETL is reached. Additionally, a direct redundancy LLR (DRLLR) is computed following fault detection to provide false alarm protection. The process observed is the duplex signal difference, and the LLR mean is of BFM magnitude with sign consistent with the detected failure. When this LLR crosses a positive threshold a false alarm is declared, the isolation process ceases, and the detection process is reinitiated. Finally, signal selftest is continually applied to all signals having unfailed status.

3.2 Rotational Kinematics

Rotational kinematics (RK) is used for fault isolation in the rate gyros and attitude platforms. The roll, pitch, and yaw rate gyros provide measurements of the aircraft body rates p, q, and r about the aircraft x, y, and z axes, respectively; and these body rates are related to the rates of change of the Euler angles measured by the attitude platforms. Thus, following a rate gyro fault detection at time t_1, the RK residual for instrument j of the suspect type is calculated at general time t_k using the appropriate equation from the following:

$$\gamma_k(p^j) = \sum_{i=1}^{k} \left\{ \bar{p}_i^j T - [\phi_i - \phi_{i-1} - (\psi_i - \psi_{i-1}) \overline{\sin \theta_i}] \right\} \tag{3}$$

$$\gamma_k(q^j) = \sum_{i=1}^{k} \left\{ \bar{q}_i^j T - [(\theta_i - \theta_{i-1}) \overline{\cos \phi_i} + (\psi_i - \psi_{i-1}) \overline{\cos \theta_i} \; \overline{\sin \phi_i}] \right\} \tag{4}$$

$$\gamma_k(r^j) = \sum_{i=1}^{k} \left\{ \bar{r}_i^j T - [-(\theta_i - \theta_{i-1}) \overline{\sin \phi_i} + (\psi_i - \psi_{i-1}) \overline{\cos \theta_i} \; \overline{\cos \phi_i}] \right\} \tag{5}$$

In Eqs. (3) - (5) and all subsequent equations, T is the ARM sample period and an overbar indicates that the quantity represents the average of its present and previous sample values. This averaging is used to reduce computational errors during high angular rate maneuvers, and the forms of Eqs. (3) - (5) avoid differentiation of the noisy attitude measurements.

At every sample time t_k, following fault detection, the residual for each suspect rate gyro is used to update its MLLR using Eq. (1), where the mean has magnitude equal to the rate gyro BFM times $(t_k - t_0)$, the variance reflects attitude gyro noise variance, and the worst-case error is computed as the sum of the magnitudes of terms reflecting initial attitude gyro noise, attitude gyro bias, rate gyro misalignment and rate gyro scale factor error. The residuals given by Eqs. (3) - (5) are also used in low-pass filters, at shallow roll and pitch angles, to update bias estimates for each unfailed rate gyro.

Treating the roll, pitch, and yaw attitude signals from a single platform as independent signal types, the RK residual for signal j of an attitude type (ϕ, θ, ψ) having a detected fault is calculated at time t_k as follows:

$$\gamma_k(\phi^j) = \sum_{i=1}^{k} \left\{ \phi_i^j - \phi_{i-1}^j - [\bar{p}_i T + (\psi_i - \psi_{i-1}) \overline{\sin \theta_i}] \right\} \tag{6}$$

$$\gamma_k(\theta^j) = \sum_{i=1}^{k} \left\{ \theta_i^j - \theta_{i-1}^j - T [\bar{q}_i \overline{\cos \phi_i} - \bar{r}_i \overline{\sin \phi_i}] \right\} \tag{7}$$

$$\gamma_k(\psi^j) = \sum_{i=1}^{k} \left\{ \psi_i^j - \psi_{i-1}^j - [(\phi_i - \phi_{i-1} - \bar{p}_i T) \cdot \overline{\sin \theta}_i + T(\bar{q}_i \overline{\sin \phi}_i + \bar{r}_i \overline{\cos \phi}_i) \cdot \cos \theta_1] \right\} \tag{8}$$

In order to calculate the attitude gyro MLLRs, it is necessary to have stored the instantaneous RK residuals, corresponding to the terms enclosed in braces in Eq. (6)-(8), in a moving window of the same length as the detector window for each signal type. At the

time of fault detection, the instantaneous residual window for each suspect signal is processed using the appropriate equation from Eq. (6)-(8) to form the MLLR residual for each window time, where t_1 corresponds to the time associated with the oldest window element. The MLLRs are computed at each intermediate window time using Eq. (1), where the mean has BFM magnitude, the variance is of the order of the attitude gyro noise variance, and the worst-case error is the sum of terms reflecting initial attitude gyro error, rate gyro bias, rate gyro misalignment and roll rate gyro scale factor error. After processing the entire window of instantaneous residuals and calculating the MLLRs corresponding to the present time, the threshold logic of Eq. (2) is applied to the lower MLLR. If no unconditional failure declaration is made, the detected failure indication is removed and the detection process proceeds smoothly on the next sample.

We note that the framework used for the attitude gyros of processing a stored window of instantaneous residuals before applying the MSPRT threshold logic is also utilized for the Mach meters and altimeters, and these five signal types do not require direct redundancy LLR calculations since the isolation process does not extend beyond the time of detection. Additionally, the MLLR for any signal of these five types is reset to zero whenever it becomes positive. This is done to accommodate the uncertainty in failure time within the window, and requires a slightly higher magnitude threshold for the MSPRTs for these sensor types.[9] For the 10^{-4} misclassification probabilities used in this study, the threshold δ in Eq. (2) is -11.4 for these five signal types and -9.2 for the others.

3.3 Altitude Kinematics

Altitude kinematics (AK) refers to the redundancy between changes in altitude measured by the altimeters and changes in altitude computed from the vertical acceleration measured by the accelerometers. This redundancy is utilized in isolating failures in the altimeters, normal accelerometers and lateral accelerometers.

The vertical acceleration, as measured by the accelerometers, is computed at time t_i as

$$Av_i = Ax_i \overline{\sin \theta_i} - (Ay_i \overline{\sin \phi_i} + Az_i \overline{\cos \phi_i}) \overline{\cos \theta_i} - g \qquad (9)$$

In Eq. (9) and elsewhere, g is the acceleration of gravity and Ax, Ay, and Az represent the actual accelerometer readings plus angular rate and angular acceleration correction terms to yield the acceleration at the center of mass. A third-order discrete filter is implemented for each altimeter to estimate altitude and vertical velocity, driven by the vertical acceleration of Eq. (9) and incorporating each altimeter's output as a measurement update on every ARM sample. (A third-order filter is necessitated by the first-order lag in the barometric altimeters.) The nonzero measurement gains are chosen to keep the effects of accelerometer bias and altimeter quantization small, and a filter trap logic is utilized to accommodate altimeter hysteresis. Each altimeter's measurement residual, the difference between its output and its propagated altitude estimate before measurement incorporation, is stored in a moving window.

At the time of altimeter fault detection, each sensor's window of residuals is processed to convert them to zero-gain residuals in order to enhance the failure signature[4], i.e., the resulting residuals are those that would have ensued if the filters were run open loop (without measurement incorporation) from the time corresponding to the oldest window element to the present time. These instantaneous zero-gain residuals for each altimeter are then summed in a manner analogous to Eq. (6) - (8) in order to form each altimeter's AK residual at the intermediate window times, and these residuals are used via Eq. (1) to form the altimeter MLLRs from the oldest window time to the present time. The threshold logic of Eq. (2) is then applied to the lower MLLR. In the MLLR calculations, the mean has altimeter BFM magnitude, the variance reflects altimeter noise variance, and the worst-case error is the sum of the magnitudes of terms reflecting initial altimeter error, initial vertical velocity estimation error, and accelerometer bias.

Following the detection of a lateral or normal accelerometer fault, two versions of Eq. (9) are integrated, each using the output of a different sensor of the suspect type. The difference between these integrals and the measured altitude at each succeeding sample forms the AK residual for each suspect sensor. The AK MLLR mean is obtained by integrating the appropriate term of Eq. (9) with the suspect sensor BFM replacing the acceleration. The worst-case error is computed as the sum of the magnitudes of terms reflecting initial altimeter error, initial vertical velocity estimation error, transonic altimeter behavior, and altimeter error due to angle of attack change, plus a term reflecting normal accelerometer scale factor error in the normal accelerometer MLLRs and lateral accelerometer misalignment in the lateral accelerometer MLLRs.

3.4 Translational Kinematics

Translational kinematics (TK) refers to the redundancy between changes in aircraft air-relative velocity measured by the air data sensors and changes in aircraft velocity obtained by integrating the acceleration computed using inertial sensor outputs. The ARM algorithm utilizes TK to isolate faults in the longitudinal accelerometers, normal accelerometers, and Mach meters.

The TK residual for longitudinal accelerometer j at time t_k, following a detected fault at time t_1, is given by

$$\gamma_k(Ax^j) = \sum_{i=1}^{k} T\left[Ax_i^j - \overline{\sin\theta_i}\, g + Vs_i M_i (r_i \sin\beta_i - q_i \sin\alpha_i)\right] - \left[Vs_k M_k \cos\alpha_k - Vs_0 M_0 \cos\alpha_0\right] \quad (10)$$

where Vs is the speed of sound, periodically recomputed as a function of altitude, and α, β, and M are the measured angle of attack, sideslip angle, and Mach number, respectively. (The measured angle of attack and sideslip angle are the vane readings compensated for aircraft pitch and yaw rates and supersonic alpha vane bias. The measured Mach number is the sensor output compensated for bias.) In the MLLR calculations, the mean at time t_k has magnitude equal to the longitudinal accelerometer BFM times $(t_k - t_0)$, the variance reflects air data sensor noise, and the worst-case error is the sum of the magnitudes of terms reflecting initial air data noise, a wind-shear doublet, misalignment of the suspect accelerometer, and transonic Mach meter behavior.

The TK residual for normal accelerometer j at time t_k, following fault detection at time t_1, is given by

$$\gamma_k(Az^j) = \sum_{i=1}^{k} T\left[Az_i^j - \overline{\cos\theta_i}\;\overline{\cos\phi_i}\, g + Vs_i M_i (q_i \cos\alpha_i - p_i \sin\beta_i)\right] - \left[Vs_k M_k \sin\alpha_k - Vs_0 M_0 \sin\alpha_0\right] \quad (11)$$

The MLLR mean has magnitude equal to the normal accelerometer BFM times $(t_k - t_0)$, the variance reflects air data sensor noise, and the worst-case error is the sum of the magnitudes of terms reflecting initial air data noise, a wind-shear doublet, suspect accelerometer scale factor error, and pitch rate gyro bias. The wind-shear doublet magnitude and MLLR variance for the longitudinal and normal accelerometers each assume one of two values depending upon the binary output of a wind turbulence filter operating on lateral channel TK residuals.

The TK residual for Mach meter j is calculated at time t_k as

$$\gamma_k(M^j) = \sum_{i=1}^{k} \left\{ Vs_i M_i^j \cos\alpha_i - Vs_{i-1} M_{i-1}^j \cos\alpha_{i-1} - T\left[Ax_i - \overline{\sin\theta_i}\, g + Vs_i M_i^j (r_i \sin\beta_i - q_i \sin\alpha_i)\right] \right\} \quad (12)$$

As for the attitude gyros and altimeters discussed earlier, the instantaneous residuals for each Mach meter, the terms in braces in Eq. (12), are stored in a moving window. At the time a Mach meter fault is detected, each sensor's instantaneous residual window is processed using Eq. (12) to compute its TK residual at the intermediate window times, and simultaneously Eq. (1) is used to compute each Mach meter's MLLR, with the threshold logic of Eq. (2) applied to the lower MLLR following complete window processing. The MLLR mean has magnitude equal to the Mach meter BFM times (Vs cos α) at the time of detection, the variance reflects the effect of Mach meter noise, and the worst-case error is the sum of terms arising from initial Mach meter error and acceleration uncertainty.

3.5 Translational Dynamics

Translational dynamics (TD) refers to the redundancy between the acceleration of the aircraft measured by the accelerometers and the acceleration predicted by stored aerodynamic coefficient functions using air data sensor measurements. TD residuals are used by the ARM algorithm to isolate failures in the lateral accelerometers and alpha vanes.

The TD residual for lateral accelerometer j at time t_k is given by

$$\gamma_k(Ay^j) = Ay_k^j - (CYB_k \beta_k + CYDR_k R_k) Q_k S/m_k \quad (13)$$

where CYB and CYDR are stored lateral coefficient functions of Mach and alpha, R is measured rudder position, Q is computed dynamic pressure, S is the surface area of the wing, and m is the estimated aircraft mass. On every sample following fault detection, the residuals given by Eq. (13) are used in Eq. (1) to compute the lateral accelerometer TD MLLRs. The TD MLLR mean has lateral accelerometer BFM magnitude, the variance reflects air data noise and the worst-case error is the sum of the magnitudes of terms reflecting the effects of beta vane bias, lateral accelerometer misalignment, neglected lateral coefficients, and scale factor error in the computed aerodynamic sideforce.

The TD residual for alpha vane j at time t_k is given by

$$\gamma_k(\alpha^j) = -(L_k^j \cos \alpha_k^j + D_k^j \sin \alpha_k^j)/m_k - Az_k \qquad (14)$$

where the lift, L, and drag, D, are computed using each alpha vane output individually in stored functions of Mach, elevator position, and angle of attack. The alpha vane TD MLLR mean has magnitude equal to the alpha vane BFM times the magnitude of the computed TD residual gradient, the variance assumes one of two values depending upon indicated turbulence level, and the worst-case error is the sum of the magnitudes of terms reflecting the effects of normal accelerometer scale factor error and aerodynamic coefficient error.

4. FLIGHT TEST RESULTS

The ARM algorithm was implemented in the flight control computers of the F-8 Digital Fly-by-Wire research aircraft in a parallel mode. Thus, ARM results were not used by the F-8 primary flight control system software. Instead, the performance and operation of the ARM software was monitored by resident software. The ARM algorithm used only two sensors of each sensor set, thus simulating a duplex sensor set. The ARM software includes extensive error simulation and signal fault insertion capability, controlled by the pilot through the computer input panel (CIP), to allow inflight evaluation of ARM algorithm performance. Signal faults that can be simulated include bias, drift, scale-factor error, hardover, transient pulse, and loss of signal. Errors that can be simulated include scale-factor errors in the lift and side-force aerodynamic coefficients, and misalignment of one triad of rate gyros or accelerometers. In addition, the choice of which of three stored aircraft mass values is used in the ARM TD calculations is updated by the pilot, on the basis of fuel remaining, via CIP entry.

The Phase I ARM algorithm flight test program involved the initial version of the algorithm, which had been designed using telemetry data from preliminary aircraft flight tests during moderate maneuvers. During this initial testing, the ARM algorithm performed well when confronted with two actual inflight failure situations.

In one instance, an opening developed in the potentiometer for alpha vane two. The ARM algorithm declared alpha vane two unconditionally failed twenty-four seconds before the baseline code, which by itself cannot isolate the failed sensor of a duplex set, detected the alpha vane fault. ARM algorithm performance during this inflight alpha vane fault isolation is shown in Figure 1. At that time the aircraft was at an altitude of 12 km and Mach 1.05. The first frame of the figure shows a normal acceleration of approximately -0.9 g. The second frame shows the two alpha vane readings. The third frame shows the TD residuals for the two alpha vanes and the TD worst-case error magnitude. The fourth frame shows the TD MLLR for each vane, indicating alpha vane two's conditional failure declaration at the time of detection and unconditional failure declaration twenty-four samples later. The positive jump in each of the MLLRs at sample thirteen, delaying the isolation process slightly, is a result of a spike in the worst-case error magnitude at that sample. Simply stated, the calculated worst-case error contribution from aerodynamic coefficient error is relatively low within a stipulated range of average vane readings and relatively high outside that region. The spike in the output of alpha vane two at sample thirteen drives the average of the vane readings outside the "good fit" region, resulting in the spike in worst-case error magnitude. The alpha interval defining the good fit region was found to be extremely conservative during Phase I flight testing and has been widened in the Phase II algorithm.

The second inflight malfunction occurred when the baseline system detected the memory parity failure of one of the triplex computers and declared it failed. Because the failed computer was dedicated to reading the number one sensor outputs and because of the mechanization of the buffer refreshing operation, the computer loss manifested itself to the ARM algorithm as all number one signals jumping to their negative maximum values. ARM sensor output selftest immediately declared all number one signals provisionally failed, and the analytic redundancy tests declared all number one signals unconditionally failed within 1.3 seconds of the computer failure.

Because the Phase I ARM algorithm had been designed using telemetry data representative of only moderate maneuvers, it was anticipated that the Phase I ARM flight tests might reveal deficiencies in the algorithm during extreme maneuvers. This was indeed the case, but the only major deficiency revealed was the inability of the relatively simple air data sensor models employed (in particular, those for angle of attack and altitude) to adequately reflect the effects of extreme maneuvers on the airflow and pressure distributions around these instruments. The uncertainty associated with the exact nature of these effects precluded their elimination through more complex sensor models, although a first-

order altimeter lag and compensation terms for vane error due to body rotational rates were added. Thus, for the most part, the air data sensor model deficiencies have been accommodated in the Phase II algorithm, outlined in the previous sections, through increases in the existing worst-case error coefficients and, in some cases, through an additional term in the TD, TK, and AK tests.

In the Phase II portion of the flight test program, the vertical and directional gyros were replaced with upgraded all-attitude platforms to alleviate the problems attributed to the attitude instruments in the Phase I flight test program. The flight test results described below reflect the Phase II algorithm performance using the all-attitude platforms.

Table 1 indicates the BFM sizes and detection window lengths for the Phase II algorithm. The ARM sample period, T, is three control law sample periods (60 milliseconds). Table 2 summarizes the average times between software fault insertion and fault isolation with the aircraft at trim at an altitude of 6.1 km and a Mach number of 0.6, and these times are in close agreement with predicted values. Sensor output selftest provisionally failed all sensors with hardover faults immediately. The isolation times shown in Table 2 also reflect ARM algorithm behavior during moderate maneuvers.

Figures 2 through 7 give some representative results of the ARM algorithm's performance in isolating simulated faults during maneuvers, both with and without simulated instrument alignment errors and aerodynamic coefficient errors. Each of the figures has four frames, with one frame showing the two signals of the failed type as seen by the ARM algorithm, one frame showing the two residuals and the worst-case error magnitude for the relevant analytic redundancy test, and one frame showing the two MLLRs for that test. The remaining frame in Figures 3 through 7 indicates the ARM algorithm estimate of a variable of interest during the test, while in Figure 2 the remaining frame indicates the DRLLR for the failed signal type.

Figure 2 indicates isolation of a simulated 0.5 scale factor in pitch rate gyro one. During the early negative pitch rate maneuver, the failure is detected but it cannot be isolated before the magnitude of the pitch rate decreases to the extent that the difference between the two pitch rate signals becomes insignificant. This forces the DRLLR to exceed the +9.2 threshold, removing the detected failure flag and reinitializing the detection process. (With the detected failure flag removed, all variables associated with fault isolation remain unchanged.) During the subsequent positive pitch rate maneuver, the fault is again detected and it is isolated when the MLLR for instrument one falls below the threshold of -9.2.

Figure 3 shows the isolation of a simulated 0.053 rad/s bias in pitch rate gyro one. This isolation is accomplished during the simulated misalignment of the number two triad of rate gyros by 0.02 rad with respect to each aircraft axis, with the aircraft rolling at a rate of approximately 1.15 rad/s. The effect of the simulated misalignment can be seen as a negative ramp in the RK residual for unfailed pitch rate gyro two. However, the magnitude of this residual is always less than the worst-case error magnitude, and pitch rate gyro one is declared unconditionally failed twenty-two samples after detection.

Figure 4 shows the isolation of a simulated -0.3 g bias in lateral accelerometer one during a wings-level sideslip of approximately 0.05 rad at a dynamic pressure of approximately 22.5 kPa (420 lb/ft^2). Because the AK test for the lateral accelerometers is effective only at high bank angles, only the TD test will be discussed. During this time, the TD test is using a lateral aerodynamic coefficient 0.9 times the value calculated from the stored polynomials in order to simulate error in the knowledge of this parameter. The effect of this simulated coefficient error can be seen as a negative bias, smaller in magnitude than the total worst-case error, in the residual for unfailed lateral accelerometer two. The ARM algorithm correctly isolates lateral accelerometer one as unconditionally failed, using the TD test, nine samples after failure simulation begins.

Figure 5 shows the successful isolation of a simulated 0.053 rad bias in alpha vane one, with the TD calculations utilizing a value for the lift coefficient 0.9 times the computed value. The aircraft is executing a 3.4 g turn at Mach 0.83 at an altitude of 5.4 km. The effect of the erroneous lift coefficient is reflected in residuals for the two alpha vanes that are more positive than they would normally be, and this can be seen as a positive bias in the residual for unfailed alpha vane two. Although this residual has a sign consistent with the sign of the detected failure, it is smaller than the postulated worst-case error magnitude. Thus, the MLLR for alpha vane two remains positive, while alpha vane one is declared conditionally failed five samples after failure insertion and is declared unconditionally failed five samples later.

Although this example shows correct alpha vane fault isolation by the ARM algorithm in spite of a ten percent error in calculated lift coefficient, this is not always the case throughout the aircraft's flight envelope. However, although some situations have been observed during 1.5 BFM bias insertions in which the ARM algorithm is unable to decide which alpha vane has failed, no instances of misisolation of the unfailed vane have been encountered when using ten percent lift coefficient error.

Figure 6 shows the successful isolation of a -0.3 g bias in longitudinal accelerometer one, with the simulated misalignment of longitudinal accelerometer two 0.02 rad about the aircraft pitch axis. The failure is inserted just before the aircraft begins a 2.5 g windup turn. The effect of the simulated misalignment and the windup turn can be

seen as an increasing positive TK residual for instrument two. Because this residual is significantly smaller than the postulated worst-case error, the MLLR for instrument two remains positive. Longitudinal accelerometer one is declared unconditionally failed 129 samples after failure injection. As shown in frame three of Figure 6, the postulated worst-case error is extremely conservative relative to the observed residual of the unfailed instrument, and this conservatism results in a relatively long isolation time of nearly 8 seconds. Significantly faster isolation performance could be obtained by lowering the magnitude (perhaps to zero) of the worst-case error wind-shear doublet used during those times, such as for Figure 6, when the turbulence estimate is low.

Figure 7 shows the successful isolation of a 0.3 g bias in normal accelerometer one during a 3 g turn (rendering the AK test ineffective). The TK test for the normal accelerometers is very sensitive to errors in the knowledge of the pitch rate, and extensive analysis of flight data indicates that pitch rate gyro two has a scale factor of approximately 0.95. The effect of this scale factor error on the estimated pitch rate during the windup turn having a pitch rate of 0.09 rad/s results in the negative ramp for the TK residual for normal accelerometer two seen in Figure 7. Since the residuals for both sensors are identically affected by the pitch rate error, the slope of the positive ramp failure signature for normal accelerometer one is smaller than it would be in nonrotating flight, accounting for the larger number of samples (268) required for isolating the failure. As for the longitudinal accelerometers, the isolation time for the normal accelerometers could be lowered by decreasing the magnitude of the low-turbulence wind-shear doublet used in the worst-case error term.

5. CONCLUSIONS

The flight tests of the analytic redundancy management algorithm formulated for the F-8 DFBW aircraft have established the validity and integrity of the approach. The method was found to be capable of isolating simulated sensor faults to the faulty sensor in a duplex set of typical flight control sensors with a high degree of confidence. In addition, two actual in-flight failure conditions were correctly handled by the algorithm. The algorithm was refined to the point where no false alarms were observed while retaining a high degree of failure isolation ability during maneuvering flight. Several conclusions are drawn from the flight research programs:

- The ARM algorithm as formulated was able to easily accommodate unforeseen sensor behavior through the worst-case error terms.
- The use of a bias failure hypothesis does not significantly restrict the ability to isolate nonbias failures in duplex systems. However, if more than one simplex sensor is involved in an analytic redundancy relationship, elaborate failure mode modeling is required for reliable FDI.
- Reasonability checks in the form of sensor output selftest are a powerful and effective complement to analytic redundancy tests.
- Analytic redundancy sensor FDI is much more sensitive to sensor modeling errors than conventional techniques. Those sensor characteristics that cannot be accurately modeled must be accommodated in order to maintain reliability of the analytic redundancy tests.
- ARM performance can be improved by estimating individual sensor biases through continuous monitoring of analytic redundancy residuals. This would allow MLLR worst-case error bias terms to be lowered, and the levels of the estimated biases themselves could be used to make isolation decisions or for trend analysis.
- Although the ARM software is more extensive than conventional FDI algorithms, it is similar in complexity to guidance and navigation calculations and poses no unusual implementation problems. Furthermore, its modularized structure lends itself to distributed processing. Although space limitations do not permit an exhaustive survey of the ARM algorithm performance during fighter-aircraft-type maneuvers, it is fair to say that the Phase II flight test results indicate that the worst-case error terms discussed earlier reflect the dominant uncompensated error sources present in the analytic redundancy residual equations, and that acceptable performance can be obtained by fine tuning the individual worst-case error coefficients.
- The fault isolation times are consistent with requirements for airplane flight control systems. The ARM algorithm could be executed more rapidly for more time-critical situations.
- The integration of analytic redundancy with a redundant flight-crucial digital flight control system remains an applications problem. The experience in the F-8 flight test program, however, suggests that this is certainly within the state-of-the-art.

6. REFERENCES

1. K.J. Szalai, P.G. Felleman, J. Gera, and R.D. Glover, "Design and Test Experience with a Triply Redundant Digital Fly-by-Wire Control System," AIAA Guidance and Control Conf., San Diego, California, August 16-18, 1976, AIAA Paper 76-1911.

2. Fault Tolerance Design and Redundancy Management Techniques, AGARD-LS-109, London, England, September 1980.

3. K.J. Szalai, et al., Digital Fly-by-Wire Flight Control Validation Experience, NASA-TM-72860, The National Aeronautics and Space Administration, Washington, D.C., December 1978.

4. M.N. Desai, J.C. Deckert, and J.J. Deyst, Jr., "Dual Sensor Failure Identification Using Analytic Redundancy", AIAA Journal of Guidance and Control, Vol. 2, No. 3, May-June, pp. 213-220.

5. A.S. Willsky, "A Survey of Design Methods for Failure Detection in Dynamic Systems", Automatica, Vol. 12, Nov. 1976, pp. 601-611.

6. A. Wald, Sequential Analysis, Dover, New York, 1973, Chap. 3.

7. A. Wald and J. Wolfowitz, "Optimum Character of the Sequential Probability Ratio Test", Annals of Math. Stat., Vol. 19, 1948, pp. 326-339.

8. J.C. Deckert, "Definition of the F-8 DFBW Aircraft Control Sensor Analytic Redundancy Management Algorithm", Report R-1178, The Charles Stark Draper Laboratory, Inc., Cambridge, Mass., August 1978, pp. 10-15.

9. T.T. Chien, An Adaptive Technique for a Redundant-Sensor Navigation System, NASA-CR-140313, The National Aeronautics and Space Administration, Washington, D.C., 1971.

7. ACKNOWLEDGMENT

This work was supported by the National Aeronautics and Space Administration Dryden Flight Research Center through Contract NAS4-2675.

Table 1. Phase II ARM parameters.

Signal Type	BFM	Detection Window Size
Mach	0.05	3
Altimeter	76.2 m	5
Angle of Attack	0.035 rad	9
Longitudinal accelerometer	0.2 g	10
Lateral accelerometer	0.2 g	10
Normal accelerometer	0.2 g	10
Roll rate	0.087 rad/s	3
Pitch rate	0.035 rad/s	2
Yaw rate	0.035 rad/s	2
Roll attitude	0.087 rad	6
Pitch attitude	0.087 rad	6
Yaw attitude	0.087 rad	6

Table 2. Inserted fault average isolation times at trim (seconds).

Signal Type	1.5 BFM Bias	BFM/s Drift	Hardover
Mach	0.06	NT	0
Altimeter	0.12	NT	0
Angle of attack	0.54	2.1	0.12
Longitudinal accelerometer	7.26	3.6	0.78
Lateral accelerometer	0.3	1.38	0
Normal accelerometer	6.51	3.84	0.36
Roll rate	0.48	1.44	0.06
Pitch rate	0.9	1.8	0.24
Yaw rate	0.9	1.8	0.24
Roll attitude	0.18	NT	0
Pitch attitude	0.12	NT	0
Yaw attitude	0.18	NT	0

NT = no test

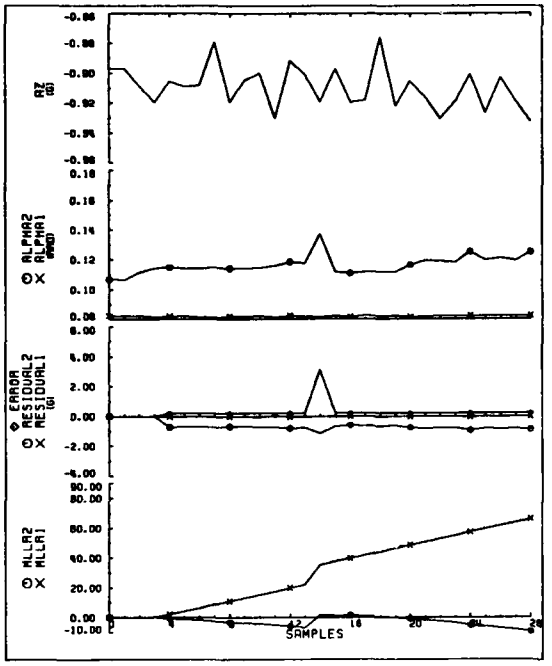


Figure 1. Actual failure of alpha vane two

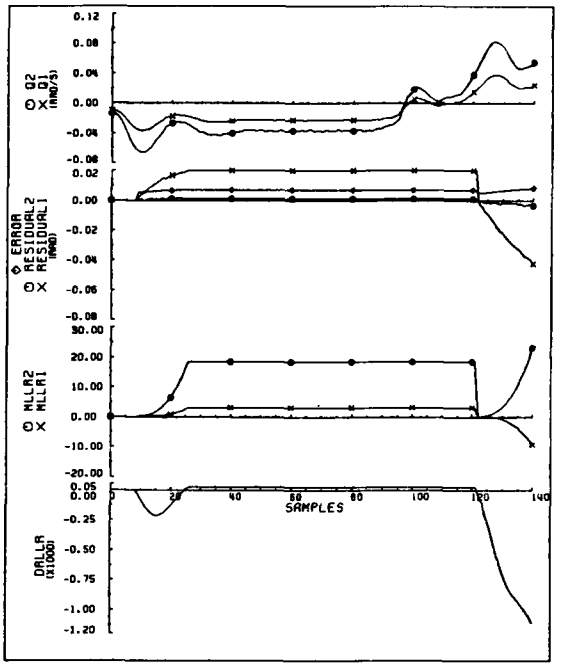


Figure 2. Pitch rate gyro one simulated 0.5 scale factor

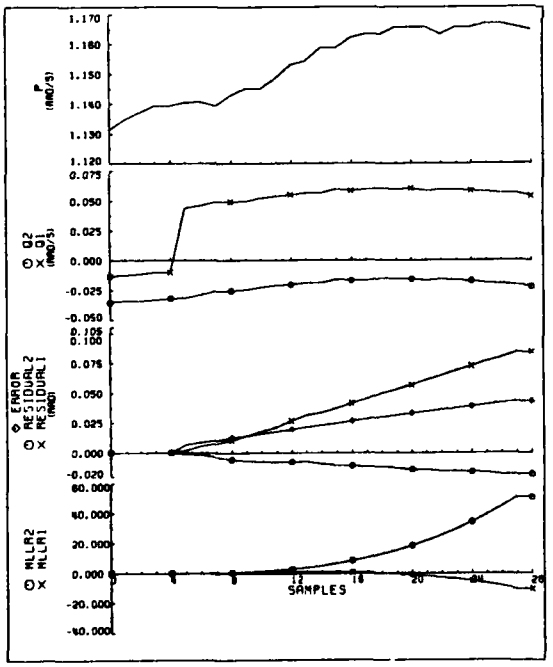


Figure 3. Pitch rate gyro one simulated 0.053 rad/s bias with simulated instrument two misalignment

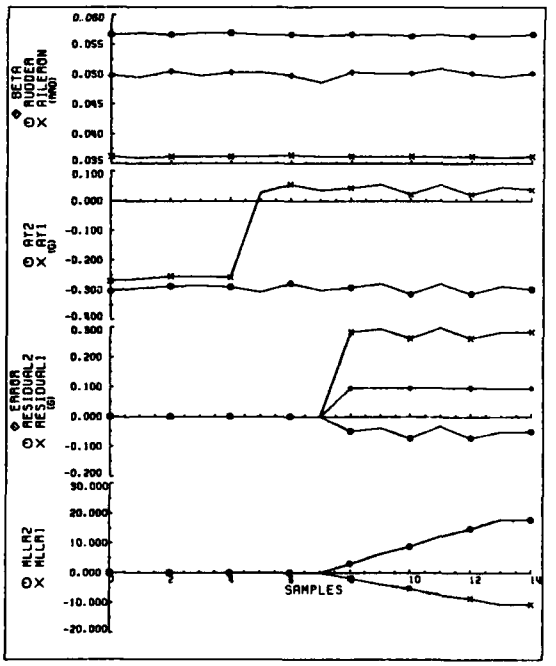


Figure 4. Lateral accelerometer one simulated 0.3 g bias using 0.9 CYB

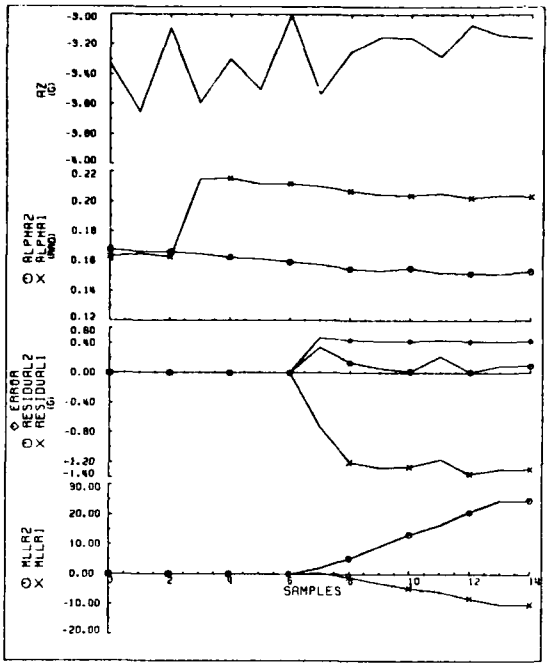


Figure 5. Alpha vane one simulated 0.053 rad bias using 0.9 CL

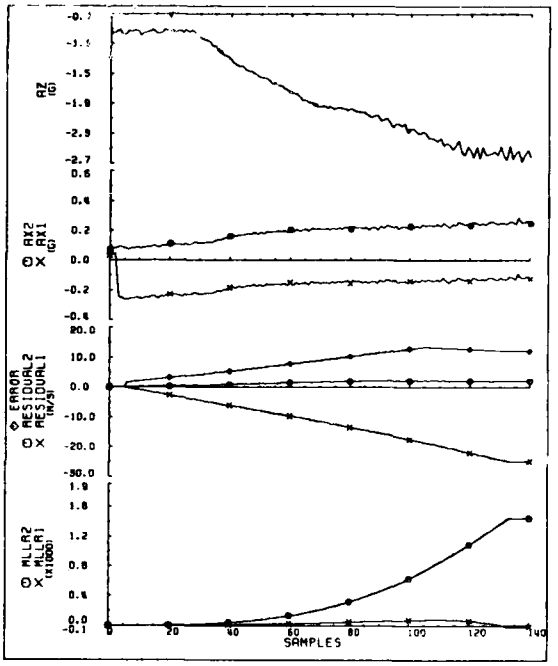


Figure 6. Longitudinal accelerometer one simulated -0.3 g bias with simulated instrument two misalignment

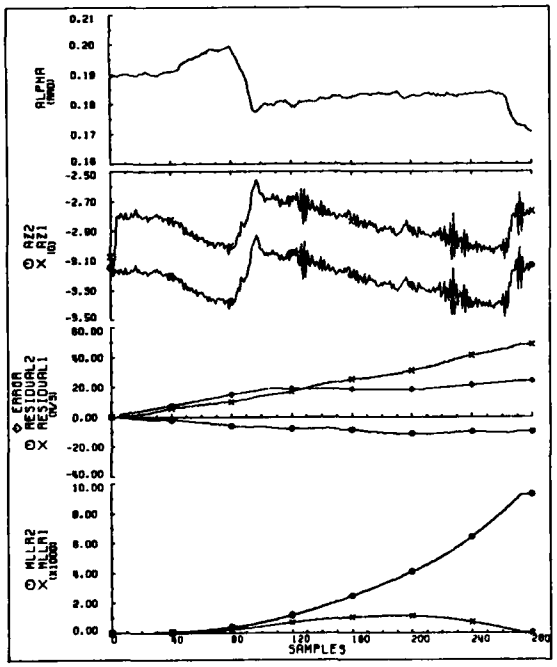


Figure 7. Normal accelerometer one simulated 0.3 g bias

INTEGRATION OF A NEW SENSOR ON TO AN EXISTING AIRCRAFT

by

A. S. LEYLAND

British Aerospace Public Limited Company
Aircraft Group
Warton Divison
Warton Aerodrome
PRESTON
Lancashire
PR4 1AX
England

INTRODUCTION

Although all sensors will have differing requirements the principles described in this paper can be applied to every integration task. To demonstrate the technique it assumed that a requirement exists to assist the crew in poor visibility by fitting a low light sensor. The requirement, and aircraft system to which it is to be fitted are fictitious, although they are representative of current systems.

The method used is Top Down Design which takes the original requirement, breaks it down to its various Functions and Processes and then recreates it into various Line Replaceable Units, and Software Requirements, in a structured computer compatible format. (See Table 1)

REQUIREMENTS

The requirements will normally be produced from an operational need, to improve reliability/maintainability or to introduce a new role. They could be in the form of introducing specified equipment which limits the scope of the design or in the form of a required improvement to the performance of the aircraft. In our case the aircrew require assistance by means of a low light sensor. We must first determine in detail, with the operator, the limits of performance expected of the new sensor system and thus establish our baseline requirements for design purposes.

We would expect to cover:-

(a) Atmospheric Conditions to define light levels, humidity, smoke or haze, temperature and precipitation in which the system is to operate.

(b) Resolution of the system, field of view, and range expected.

(c) Operational Use will need to define if the sensor is only to provide a back up to automatic flight as a pilot monitor, will it ever be used as the primary flying sensor for the pilot, do we require alternative fields of view, and is the sensor fixed, stabilized or capable of tracking. Altitude and aircraft speed need to be defined to enable the conditions, under which the sensor will be used, to be evaluated. Fall out capabilities such as navigation and fixing need to be considered.

(d) Control and Display will define the integration requirements of the sensor with the rest of the aircraft and will decide the display positioning and possible integration with a multi function display.

At this stage, as well as getting the baseline requirements defined and documented in an unambiguous form, all available documented data on similar systems and proposed sensors should be gathered and assimilated, preferably without being subjected to any sales pitches.

THE DESIGN PHASE

Having determined the size of the task in Requirements above the next stage is to plan the design phase laying out all the steps, serial and parallel, and allocate suitable personnel for each. Use must be made of appropriate specialists within the design organisation to assist particularly in areas such as:

i) Operational Analysis
ii) Flight Operations and Cockpit Ergonomics
iii) Airworthiness
iv) Quality Assurance
v) Maintainability
vi) Product support

It is also possible to contact other divisions within a large company to make use of related experience in the weapons/sensor fields. The initial design phase must consider the method by which the various activities are to be reported and how the information is to be prepared to allow specifications to be produced.

An overall plan has to be laid out so that all activities are drawn together and each individual knows what is expected of him, and the timescales for the various phases. The plan will normally consist of three parts:-

1. System Definition
2. Development
3. Production Modification

DEFINITION

Decomposition

In our Top Down Approach the first stage in definition is the decomposition of the requirements into Functional blocks. This is best done with a small number of engineers working in a 'brain-storming' environment. This session will take place in a very short timescale, measured in hours, and will define functions at the highest level of individually identifiable tasks. Functions known to exist already in the main aircraft systems must be identified as they may need some modification or may not be as useful as first thought. Although the choice of functions originally varied from team to team experience has shown that there is now a broad agreement in the functions selected for any particular task.

For a Low Light sensor the following are typical:-

F1 Scene Detection and Recognition
F2 Display
F3 Recording
F4 System Test
F5 Control
F6 I.R. Radiation Sensing
F7 Visible Sensing
F8 Radar Sensing
F9 Navigation
F10 Landing

The functions which are documented and annotated F1 - F10 are arrived at considering a mission from start to finish including all pre and post flight activities. Sensing is considered in three Functions - IR, Visible, and Radar because at this stage we have not defined the type of sensor. Detailed considerations will decide which is the optimum sensor to meet the overall requirements.

The next stage is to take each function and decompose it into the Processes carried out in that function and document this output by annotating the processes according to function e.g. F1A, F3D etc. Each function must have at least one process and usually several. At this stage many processes will be considered as the design evolves, and some discarded in an iterative process. In our study several "fall out" processes can be considered and assessed to determine whether they should be included with the original requirements.

Some of the processes to be considered are indicated below with their prime functions.

F1 Scene Detection and Recognition

This function can be considered in two main parts. The first, F1A, is the recognition of a scene covered by the sensor. This process will enable the pilot to view the scene on his display and enable him to assess the aircraft position relative to the scene presented. The requirement initially defined will tell us if the pilot has to fly using the display or if he is only to recognise ground features whilst flying using other data. Assuming a requirement to use the display to fly the aircraft then several requirements have to be defined:-

The speed range of the aircraft
The height above ground level
Atmospheric conditions

The second, F1B, part of the function is the consideration of the need to identify objects within the scene, which will require an ability to change the field of view and increase the magnification of the image.

To enable a sensor to be used as a flying aid or a monitor it is necessary to fix the sensor relative to aircraft axes. The sensor has also to be mounted in a position to give a large field of view in azimuth and elevation. The considerations of the display function F2 can place constraints on the sensor from consideration of the display geometry in terms of format, distance from the crew member's eye, the total image subtense, spatial resolution, and display range of brightness.

The following main factors must be considered:-

(a) The image size and its relationship to the outside world. A one to one correlation is necessary if the display is to be presented head up. Magnification on selected parts of the display may be required for identification of particular objects.

(b) Eye acuity will vary dependent on the cockpit environments. An assessment of the cockpit vibration at low level, high speed may need to be made.

(c) Using Johnsons criteria 2 resolvable elements are required for detection and 8 for identification of any object.

(d) Size of display area available and eye acuity will determine the field of view that can be presented.

(e) Brightness, colour, and contrast available in the display will determine the matching performance required from the sensor.

(f) Image distortion, movement, and flicker need to be considered when performance is being determined.

F2 Display

We can consider at least five process within the Display function.

F2A Display Generation must consider the form of electronic presentation to be used - raster, cursive or a mixture, and the requirements for information to be added to the scene presented by the sensor.

F2B Symbol/Alphanumeric Generation examines the requirements for flight data to be presented as an overlay to the 'outside world' display. In particular integrity of data used as a flight display must be considered.

F2C Video mixing will be required to present data from the sensor and alphanumerics from a Waveform generator. A synch system will have to be defined.

F2D Signal Processing will address the scan conversion requirements between sensor and display and will also take care of image magnification and field of view reduction.

F2E Display matching will enable the crew to determine the error in navigation by comparing the predicted position of a feature with the actual position shown on the display.

A method of driving the predicted position into coincidence with the 'true' position on the display has to be stated.

F3 Recording

A possible function for the system which could be included for post flight analysis. Possible use as a reconnaisance sensor could be included on military operations. The processes could be:

F3A Video recording
F3B Film recording

F4 Systems Test

All modern aircraft require means of testing systems either built in test (BIT) or by means of ground test equipment.

This function must define the philosophy to be applied to the new system so that the processes can be applied for each part of the system.

The tests to be performed by either interruptive or continuous BIT and the processes for failure indication have to be examined.

F5 Control

The controls can broadly be split into two sections:- those which control the sensor and those which control the presentation to the crew. It will be necessary to define the processes in this function taking account of the limitations of the existing aircraft systems such as:-

(a) Cockpit space
(b) use of existing controls particularly if an existing display is utilised.
(c) cockpit ergonimics.

It will also be necessary to consider what operations are necessary after equipment malfunctions or failures.

F6 I.R. Radiation Sensing

F7 Visible Sensing

F8 Radar Sensing

The Functions F6, F7 and F8 will in the course of a design study be normally be reduced to one sensor although a combination of say IR and Low Light Television could have advantages. In the design study the processes will be similar in each function for each type of sensor but will have differing constraints due to the use of different parts of the electro magnetic spectrum.

Factors to be considered are:-

Detector characteristics
Radiance dynamic range
Optics and field of view
Filters
Obscuration
Resolution

F8 continued

In order to provide adequate resolution the radar sensor would have to operate in the millimetric region of the spectrum.

The effect of erosion on the transparency due to sand/dust and rain needs to be considered at this stage when performance is being assessed.

F9 Navigation

The ability to identify points on the ground enables the sensor to be used for position fixing. We can consider:-

F9A On top fixing where the sensor is used to fly the aircraft over an identifiable object whose position is know.

F9B Off set fixing where a ranging sensor is pointed at an identifiable object by marking a position on the display with a fixing cross.

F10 Landing

In this function the sensor provides to the crew a view of the landing area under poor weather conditions. Normal instrument flying displays would be superimposed on the display. The system could also be used by the military on advanced airfields where normal approach lights are not available.

PROCESS ASSESSMENT

Having defined all the processes and performed the necessary iterations to arrive at a consistent suite of processes it is necessary to examine each in detail. Each will be broken down to define all its inputs and outputs which will be listed together with the source or destination.

The inputs/outputs will then be considered with the process to establish the following:

(a) Sensor obscuration
(b) Process time, sequencing and crew workload
(c) Impact of Electro Magnetic Compatibility
(d) Growth requirements or capability
(e) Reliability, availability and integrity
(f) System initialisation and shut down
(g) Ergonomic requirements
(h) Test/maintainability
(i) Cost drivers/limitations
(k) Heat dissipation/cooling
(l) Power supplies
(m) Environmental considerations
(n) Data requirements including resolution, accuracy, update rates, staleness, synchronisation, signal type.
(p) Impact on system effectiveness

System Generation

Having considered all the processes and the inputs and outputs of the data flow it is necessary to produce a system from all the information that has been documented. If we do this in a structured form it is possible to take advantage of computer aided design techniques.

A technique that has been developed for this purpose is known as CORE (Controlled Requirements Expression). In each process all the inputs and ouputs are labelled with a code giving the source or destination and the complete process is "drawn" on the computer V.D.U. including any calculations that are carried out. As all the processes are completed and filed in the computer it is possible to check all inputs and outputs of the total design for consistency. Similarly all signals will be traceable through the processes and functions.

The formality of CORE and its knowledge of all data flows assists in the definition of best way to group the design into individual Line Replaceable Units.

A further advantage is because of the formal definition of all calculations and requirements it is possible to derive directly from CORE any Software Requirements needed in the new design. Work is continuing with the aim of turning the Software Requirements derived from CORE directly into a high level language PROGRAM such as CORAL or ADA.

From the above system generation it is possible to produce Specifications including detailed input/output definition which can be used to procure the equipment which forms the new system.

Each equipment can be procured individually and qualified to its own specification and test requirements but they must be brought together to develop the system before the modification can be applied on production aircraft.

Development

The development of any system can be divided into five sections:-

1. Individual Equipment Tests
2. Sub System Testing
3. Integrated Systems Tests
4. Ground Performance Analysis
5. Flight Trials

Individual Equipment Testing

All equipments are built to a specification defining the performance to be met under various environmental conditions. Each equipment is proved in its own right against standard environmental tests and in many cases this is sufficient to predict performance and provide flight clearance. In the case of a low light sensor the performance can only be determined after exhaustive tests including some under dynamic conditions. Some tests can be made by mounting the sensor on a vehicle but before use in an operational situation tests in a specially instrumented "hack" aircraft are necessary to check the effects of aircraft manoeuvring at low level.

Use of a 'hack' aircraft allows tests to be started before the correct aircraft has been modified thus gaining time in the development of the main part of the new system.

Sub System Testing

As equipments are proved in their own right it becomes necessary to build up the sub system.

The interfaces have to be proved between all equipments and shown to be compatible in both static and dynamic conditions. Particular attention has be given to the matching of the sensor and the display particularly in the area of video processing.

Integrated Systems Testing

The new subsystem equipment must be integrated into the existing avionic system as it will be operated in the aircraft. Methods of stimulating each sensor are provided and access points used to monitor signal flows and system performance. The test rig used is built to aircraft drawings so that its first function is to check that the aircraft wiring is correct. Particular emphasis is placed on providing screening, bonding and earthing as representative as possible to the aircraft so that electro magnetic compatibility can be checked. No attempt is made to build an aircraft structure as this inhibits access to the equipments and would not be identical to the aircraft, so that EMC problems cannot be completely eliminated. Provision is made on the rig to access all wiring as necessary to measure or inject signals during operation and to assist in analysis. The sub-system will be capable of being run independently or connected into the complete system and operated by the normal cockpit controls in conjunction with rest of the avionic system. When the system has been shown to operate under static conditions on laboratory power supplies it is run on a representative aircraft distribution system from aircraft alternators. The effects of load switching, power supply failures and any other relevant conditions such as non-paralleled operation in a normally paralleled system will be assessed for malfunctions in the new sensor system.

Once static integration tests have been completed, dynamic tests are required to check the new sensor and software whilst the aircraft computers carry out their other functions. To do this a ground based computer carrying a Data Acquisition and Stimulation System (D.A.S.S.) is used. A digital computer model of the aircraft performance in aerodynamic terms which can be varied by changing the thrust, lift and drag characteristic is driven in roll, yaw and pitch. The digital model provides outputs, which are used to drive the avionic system as the sensor outputs would drive the system in an aircraft. The aircraft computer responds to these stimuli and drives the cockpit displays in a realistic manner. It is,therefore, possible to provide a video display recorded from the hack aircraft on the rig and match the flight profile by driving the D.A.S.S. with flight data recorded on the hack aircraft. We will, therefore, be able to look at the interaction between the sensor video and the symbology and alpha numerics produced on the display. After all engineering problems have been eliminated the new system can be checked by operational aircrew to determine the acceptability of the mechanisation under real time conditions.

Ground Performance Analysis

Although this process is carried out mainly during the development phase some of the tasks are started in the design phase to assist in defining some of the processes. The first task with a low light sensor is to determine the feasibility of flying an aircraft at low level using an electronic display. A computer generated display can be used in a simulator of the aircraft to be fitted with the new sensor and a selection of aircrew asked to fly the system. By varying the field of view and assessing the height at which pilots can fly at various speeds it is possible to arrive at a figure for the minimum field of view acceptable.

The second task carried out in the design phase is simulating the display presentation to the aircrew with all the symbology and alpha numerics. Using a general purpose waveform generator, various combinations of symbology may be tried until the optimum is determined by real time operation in a cockpit simulator 'flown' by aircrew.

Ground Performance Analysis continued

This display simulation has proved to be particularly cost effective as most proposals can be evaluated before aircraft computer software is written.

During the development phase the non aircraft hardware will be replaced by aircraft equipment to show the results obtained during initial simulation are still valid.

As one of the possible options of use of a low light sensor is in the landing phase it is necessary to check that the software driving the displayed data matches the integrity requirements laid down for the hardware. To do this the software has to be subjected to a theoretical and practical validation check. The use of a structured definition of software requirement enables the software to be produced in a form to aid validation. Use of a modular structure, and limiting the number of instructions in each module, allows a theoretical assessment to be made of the software performance as various parameters are changed. It will also be possible to predict from the computer generated software requirements the expected effect of all parameter changes and combination of changes. We can, therefore, perform closed loop checks on the software by inputing known parameters from a ground based computer and measuring the output from the processor containing the airborne software.

Flight Trials

Under this heading come all tests on the aircraft to prove the new system and offer a release for in service clearance.

The first testing to be carried out will be on the new transparency and fairing which have to be fitted to a part of the aircraft offering a good forward view. Wind tunnel tests will have predicted the airflow pattern round the sensor but only flight tests can confirm acceptability. These tests can be carried out with a mock-up structure fitted to the aircraft long before an actual fabricated production unit is ready. This will enable any changes to be incorporated easily and can clear any possible problems of interaction with other external loads, disturbed flow into the engines, on-set of buffet or interference with air data sensors.

Further problems to be assessed at this stage are contamination during take off or landing and possible erosion effects due to rain or sand and dust. Resistance to birdstrike will have to be done on a correct structural representation of the fairing.

When the developed system is offered for flight testing the first phase will cover all functions on the ground to confirm correct operation. Next with engines running a full check of electro magnetic compatibility will be carried out to show that operations of other on-board systems do not affect the picture quality.

A series of flights are planned to function the system over its defined flight envelope. These will take into account the nature of the new sensor and will demand a cautious approach to limiting conditions.

Normally a safety pilot will be used in the aircraft to monitor the aircraft position whilst the test pilot flies on the new sensor and associated display. The first flights will be in daylight and clear weather to assess operability over a range of heights and speeds gradually approaching the limiting condition. Finally the trial will move into low light conditions starting with full moon and no cloud before testing down to the limiting performance conditions of weather, height and speed.

During the trials it will be necessary to fly over towns at night to check the effects of glare and haze, and also to determine the effect of flying through occasional rain showers. The effect of salt accumulation due to prolonged flight over the sea has to be assessed and it may be necessary to fit a wash system to the transparency.

After checking the operation of the sensor, some systems proving flying will be necessary to assess tha ability of the crew to do on top or off set fixing and to determine the accuracy of the updates obtained. The usefulness of the sensor as a landing aid can be checked on an opportunity basis at the end of each flight without the need to schedule special flights.

Production Modification

The introduction of a new system into a production aircraft must be carried out under a modification procedure. Three areas will be involved in the case considered:-

(a) Introduction of fixed fittings and wiring
(b) Introduction of equipment
(c) Introduction of software

The fixed fittings and wiring will have to be designed for fitting on the production line for new aircraft and as a retrospective embodiment for aircraft already in service. although both methods achieve the same aim they will entail different modifications. Consideration will also need to be given to the use of the low light device as an optional fit for which the sensor and fairing can be removed if required and the aircraft flown clean. New equipment can be procured by normal methods and planned to arrive for introduction in modification kits but for any changed equipment a 'pool' must be established to enable units to be returned to manufacturers for modification.

Production Modification continued

Software can be prepared off line and as it only takes a short time to load it does not present a problem and can be loaded as the new equipment is fitted.

At the same time that a modification is introduced it will be necessary to update all the descriptive manuals, test procedures, flight operation manuals and parts lists for spares holding. The publications etc., can be as big a task as the rest of the modification and too often are not started in time at either the equipment manufacturer or the aircraft company.

Finally provision must be made for all the necessary ground support equipment needed to service and maintain the new system including, in our case, the possible introduction of optical test benches and special sources of radiation.

The above although not being definitive on the modification does indicate the steps necessary to introduce a new sensor into an aircraft which is already fitted with a comprehensive avionic system.

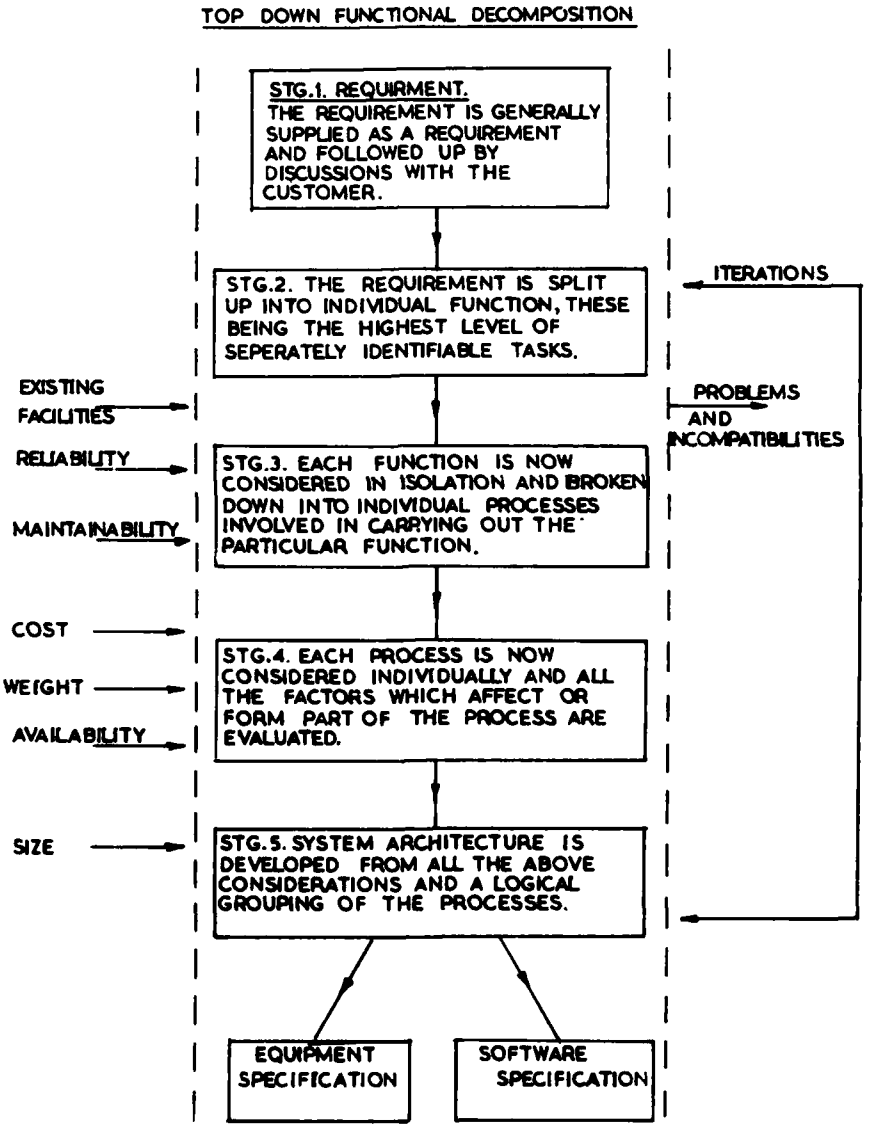


TABLE.1.

REPORT DOCUMENTATION PAGE			
1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-GCP/AG-272	ISBN 92-835-1451-3	UNCLASSIFIED

5. Originator	Advisory Group for Aerospace Research and Development North Atlantic Treaty Organization 7 rue Ancelle, 92200 Neuilly sur Seine, France
6. Title	ADVANCEMENTS IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS
7. Presented at	
8. Author(s)/Editor(s)	Edited by: Mr J.L.Hollington
9. Date	June 1983
10. Author's/Editor's Address	Smiths Industries, Aerospace and Defence Systems Company Cheltenham Division, Bishops Cleeve, Cheltenham, Glos GL53 4SF, UK
11. Pages	152
12. Distribution Statement	This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.

13. Keywords/Descriptors

E/O sensors and airborne radar
New types of inertial sensor
Low cost fluidic sensors
Measurement of airspeed and windshear with an airborne laser
Analytical redundancy and integration of new sensor on existing aircraft
Speech recognition

14. Abstract

The purpose of this AGARDograph is to provide the general reader with info on advances in the field of sensors and current approaches to problems of integration into aircraft G & C systems. Individual papers on: advances in E/O sensors and airborne-radar, New types of inertial sensor, Low cost fluidic sensors, Measurements of airspeed and windshear with an airborne laser, Application of analytical redundancy and integration of a new sensor on to an existing aircraft. One paper on automatic speech recognition is included.

Prepared at the request of the Guidance and Control Panel of AGARD.

AGARDograph No.272 Advisory Group for Aerospace Research and Development, NATO ADVANCEMENTS IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS by Mr J.L.Hollington Published June 1983 152 pages The purpose of this AGARDograph is to provide the general reader with info on advances in the field of sensors and current approaches to problems of integration into aircraft G & C systems. Individual papers on: advances in E/O sensors and airborne-radar, New types of inertial sensor, Low cost fluidic sensors, Measure- P.T.O	AGARD-AG-272 E/O sensors and airborne radar New types of inertial sensor Low cost fluidic sensors Measurement of airspeed and windshear with an airborne laser Analytical redundancy and integration of new sensor on existing aircraft Speech recognition	AGARDograph No.272 Advisory Group for Aerospace Research and Development, NATO ADVANCEMENTS IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS by Mr J.L.Hollington Published June 1983 152 pages The purpose of this AGARDograph is to provide the general reader with info on advances in the field of sensors and current approaches to problems of integration into aircraft G & C systems. Individual papers on: advances in E/O sensors and airborne-radar, New types of inertial sensor, Low cost fluidic sensors, Measure- P.T.O	AGARD-AG-272 E/O sensors and airborne radar New types of inertial sensor Low cost fluidic sensors Measurement of airspeed and windshear with an airborne laser Analytical redundancy and integration of new sensor on existing aircraft Speech recognition
AGARDograph No.272 Advisory Group for Aerospace Research and Development, NATO ADVANCEMENTS IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS by Mr J.L.Hollington Published June 1983 152 pages The purpose of this AGARDograph is to provide the general reader with info on advances in the field of sensors and current approaches to problems of integration into aircraft G & C systems. Individual papers on: advances in E/O sensors and airborne-radar, New types of inertial sensor, Low cost fluidic sensors, Measure- P.T.O	AGARD-AG-272 E/O sensors and airborne radar New types of inertial sensor Low cost fluidic sensors Measurement of airspeed and windshear with an airborne laser Analytical redundancy and integration of new sensor on existing aircraft Speech recognition	AGARDograph No.272 Advisory Group for Aerospace Research and Development, NATO ADVANCEMENTS IN SENSORS AND THEIR INTEGRATION INTO AIRCRAFT GUIDANCE AND CONTROL SYSTEMS by Mr J.L.Hollington Published June 1983 152 pages The purpose of this AGARDograph is to provide the general reader with info on advances in the field of sensors and current approaches to problems of integration into aircraft G & C systems. Individual papers on: advances in E/O sensors and airborne-radar, New types of inertial sensor, Low cost fluidic sensors, Measure- P.T.O	AGARD-AG-272 E/O sensors and airborne radar New types of inertial sensor Low cost fluidic sensors Measurement of airspeed and windshear with an airborne laser Analytical redundancy and integration of new sensor on existing aircraft Speech recognition

ments of airspeed and windshear with an airborne laser, Application of analytical redundancy and integration of a new sensor on to an existing aircraft; One paper on automatic speech recognition is included.

This AGARDograph was prepared at the request of the Guidance and Control Panel of AGARD.

ISBN 92-835-1451-3

ments of airspeed and windshear with an airborne laser, Application of analytical redundancy and integration of a new sensor on to an existing aircraft; One paper on automatic speech recognition is included.

This AGARDograph was prepared at the request of the Guidance and Control Panel of AGARD.

ISBN 92-835-1451-3

ments of airspeed and windshear with an airborne laser, Application of analytical redundancy and integration of a new sensor on to an existing aircraft; One paper on automatic speech recognition is included.

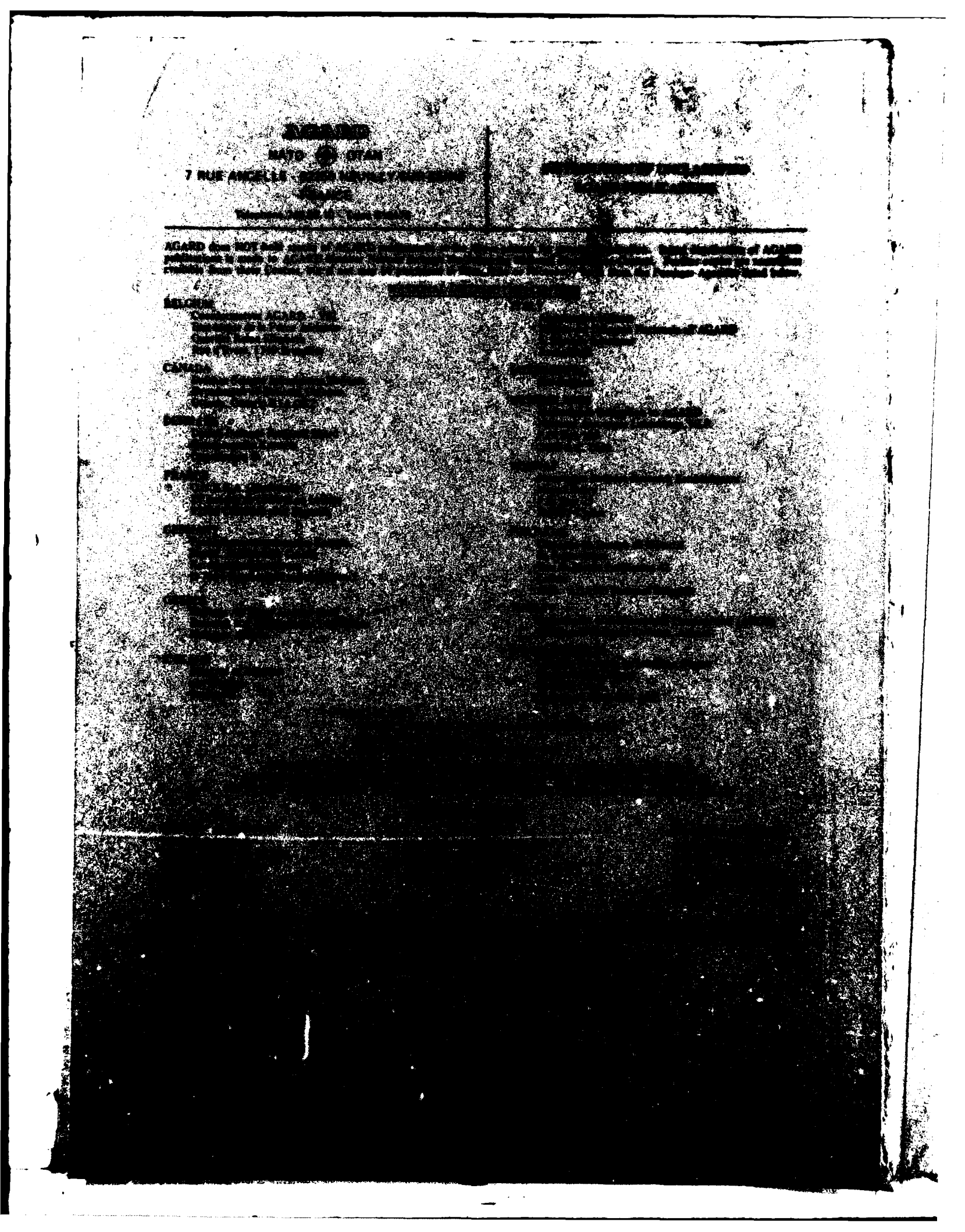
This AGARDograph was prepared at the request of the Guidance and Control Panel of AGARD.

ISBN 92-835-1451-3

ments of airspeed and windshear with an airborne laser, Application of analytical redundancy and integration of a new sensor on to an existing aircraft; One paper on automatic speech recognition is included.

This AGARDograph was prepared at the request of the Guidance and Control Panel of AGARD.

ISBN 92-835-1451-3



END

DATE
FILMED

10 83

DTIC